Electronic Journal of Statistics
Vol. 8 (2014) 2022–2065
ISSN: 1935-7524
DOI: 10.1214/14-EJS943

Nonparametric link prediction in large scale dynamic networks

Purnamrita Sarkar[*] **and Deepayan Chakrabarti**[†]

University of Texas, Austin
e-mail: purna.sarkar@austin.utexas.edu
deepayan.chakrabarti@mccombs.utexas.com

Michael Jordan

University of California, Berkeley
e-mail: jordan@cs.berkeley.edu

Abstract: We propose a nonparametric approach to link prediction in large-scale dynamic networks. Our model uses graph-based features of pairs of nodes as well as those of their *local neighborhoods* to predict whether those nodes will be linked at each time step. The model allows for different types of evolution in different parts of the graph (e.g, growing or shrinking communities). We focus on large-scale graphs and present an implementation of our model that makes use of locality-sensitive hashing to allow it to be scaled to large problems. Experiments with simulated data as well as five real-world dynamic graphs show that we outperform the state of the art, especially when sharp fluctuations or nonlinearities are present. We also establish theoretical properties of our estimator, in particular consistency and weak convergence, the latter making use of an elaboration of Stein's method for dependency graphs.

MSC 2010 subject classifications: Primary 62G08; secondary 91D30.
Keywords and phrases: Link prediction, dynamic networks, nonparametric.

Received April 2013.

Contents

[*] This work was partly done when the author was at the University of California, Berkeley.
[†] This work was partly done when the author was at Yahoo! Research.

1. Introduction

Many real-world problem domains generate data in the form of graphs or networks. Examples include social networks (e.g., Facebook), recommendation services (e.g., Netflix or Last.fm), biochemical networks, citation graphs and market analysis. The inferential problem in these settings is often one of *link prediction*. This problem can be formulated in a static setting where one assumes that a fixed but unknown graph is partially observed, and one wishes to assess whether a pair of nodes that are not known to be linked are in fact linked, given an observed linkage pattern among other nodes. Many real-world graphs are often best modeled, however, as dynamic entities, where links can arise and disappear over time. In the dynamic setting the link prediction problem involves assessing whether two nodes will be linked at time t given the linkage patterns at all previous times.

Real-world graphs of current interest are often very large, involving many hundreds of thousands or millions of nodes. The dynamic setting involves sequences of such graphs. Given the large-scale nature of these data structures, inferential methodology that may be feasible on smaller graphs of hundreds of nodes, such as Markov random fields and other graphical models, are generally infeasible for real-world link prediction problems, and practical approaches to such problems generally involve simple heuristics, such as estimating a probability of a link being present as a simple function of the last time a pair of nodes formed a link, or the number of common neighbors between a pair of nodes [14, 18, 27, 31]. While these heuristics do respect the computational imperative, and are often useful in practice, there has been little in the way of statistical analysis to provide a sound foundation for their use and to assess the quality of the inferences that they provide. This is particularly true in the dynamic setting, where link prediction is often approached by specifying various measures of connectivity in a static graph and extending these measures in an ad hoc manner to sequences of graphs.

In this paper, we develop a nonparametric methodology for link prediction in large-scale dynamic networks. Our methodology is a relatively simple kernel-based approach, one that aims to retain the virtues of the simple heuristic methods, both in their favorable computational scaling and in the relatively weak assumptions that they appear to make on the graph generation process. As compared to existing heuristic approaches, however, our kernel-based approach allows us to provide a formal inferential treatment of link prediction—we establish consistency and weak convergence of our estimator. On the computational front, while a naive implementation of a kernel method would have poor scaling (due to the need to compare query points to every point in a training set), we show that our kernel-based approach is amenable to locality sensitive hashing (LSH) [15], which provides a fast and scalable implementation of the estimator.

Our approach is in the spirit of the nonparametric autoregressive time series models [19]. In these models the evolution of a sequence x_t of continuous univariate random variables is modeled by taking the conditional expectation of x_t to be a function of a moving window $(x_{t-1}, \ldots, x_{t-p})$, and estimating this function via kernel regression. It is also possible to consider multivariate extensions of such models. While it would be possible in principle to apply such models to our problem by encoding graphs as vectors, in practice the large-scale graphs that are our focus would generate high-dimensional vector representations that would be fatal to naive kernel regression. Instead, we think of the graphs as providing a "spatial" dimension that is orthogonal to the time axis. In addition to imposing the conditional independence assumption implicit in the use of a moving window, we make the additional assumption that the linkage behavior of any node i is independent of the rest of the graph given its "local neighborhood"; in effect, local neighborhoods are to the spatial dimension what moving windows are to the time dimension.

Thus we model the out-edges of i at time t as a function of the local neighborhood of i over a moving window of time, resulting in a much more tractable problem. As a byproduct, this also allows for different evolutions for different regions to exist in the same graph; e.g., regions of slow versus fast change in links, assortative versus disassortative regions (where high-degree nodes are more/less likely to connect to other high-degree nodes), densifying versus sparsifying regions, and so on.

As a brief summary, our contributions are as follows:

(1) *Nonparametric problem formulation:* We offer, to our knowledge, the first nonparametric model for link prediction in dynamic networks. The model is powerful enough to accommodate different regions with different dynamics, which is not accommodated in existing heuristic approaches. It also allows covariates to be incorporated (such as demographic data about a node).

(2) *Consistency and weak convergence of the estimator:* We prove consistency of our estimator using notions of strong mixing in Markov chains. To establish weak convergence we show how to adapt Stein's method to our setting, going beyond the dependency graph formulation of Stein's method [26] to allow long-range weak dependence instead of marginal independence.

(3) *Fast implementation via LSH:* Nonparametric methods such as kernel regression require computing kernel similarities between a query and all members of the training set. A naive implementation would lead to computation linear

in the training set size, which is generally infeasible for large-scale networks. In order to mitigate this issue, we adapt the locality sensitive hashing algorithm of Indyk and Motwani [15] to our particular kernel function.

(4) *Empirical improvements over previous methods:* We demonstrate the empirical effectiveness of our method on link prediction tasks on both simulated and real networks. On graphs with nonlinear linkage patterns (e.g., seasonal trends), we outperform all of the state-of-the-art heuristic measures for static and dynamic graphs. This result is obtained in particular on a real-world sensor network graph. On other real-world datasets with smoother and simpler evolution, we perform as well as the best competitor. Finally, we compare our LSH-based kernel regression to exact kernel regression, and show that the LSH-based approach yields almost identical accuracy at a fraction of the computational cost.

The rest of the paper is organized as follows. We present the model and the estimator in Section 2. Our LSH implementation is described in Section 3. Section 4 provides an experimental evaluation of our method. We provide an analysis of consistency in 5. In Section 6 we discuss our adaptation of Stein's method which we use to establish weak convergence of our estimator in Section 7. We provide a discussion of related work in Section 8 and we present our conclusions in Section 9.

2. The model and the estimator

We begin by introducing some notation. Consider a sequence of directed graphs, $\mathcal{G} = \{G_1, G_2, \dots, G_t\}$. Define the indicator $Y_t(i,j)$ which equals 1 if the edge $i \to j$ exists at time t, and 0 otherwise. Let $N_t(i)$ denote the *local neighborhood* of node i in G_t; in our experiments, we define it to be the set of nodes within two hops of i and all edges between the nodes in that set (the reasoning behind this choice is explained later in Section 4.4). Note that the neighborhoods of nearby nodes can overlap. For any integer $p > 1$, let $\vec{N}_{t,p}(i) = \{N_t(i), \dots, N_{t-p+1}(i)\}$; this represents the local neighborhood of i along both spatial and temporal dimensions.

2.1. The model

Our model is as follows:

$$Y_{t+1}(i,j)|\mathcal{G} \sim \text{Bernoulli}(g(\psi_t(i,j)))$$
$$\psi_t(i,j) = \{\mathrm{s}_t\,(i,j)\,, d_t\,(i)\},$$

where $0 \le g(\cdot) \le 1$ is a function of two sets of features: those specific to the *pair* of nodes (i,j) under consideration—$\{\mathrm{s}_t\,(i,j)\}$—and those for the local neighborhood of the endpoint i—$\{d_t\,(i)\}$. We require that $\mathrm{s}_t\,(i,j)$ and $d_t\,(i)$ be functions of $\vec{N}_{t,p-1}(i)$ and $\vec{N}_{t,p}(i)$ respectively, so that $\psi_t(i,j)$ is a function of $\vec{N}_{t,p}(i)$. Thus, $Y_{t+1}(i,j)$ is assumed to be independent of $\mathcal{G}$ given $\vec{N}_{t,p}(i)$, limiting the dimensionality of the problem.

We make two observations here. First, the graphs are directed and $\psi_t(i,j)$ depends on $d_t\,(i)$ but not on $d_t\,(j)$. The intuition underlying this choice is that node i forms edges based on the behavior it observes in its neighborhood, and this is most easily expressed via directed graphs and egocentric neighborhoods around node i. Second, two pairs of nodes (i, j) and (i', j') that are close to each other in terms of graph distance are likely to have overlapping neighborhoods, and hence a higher probability of sharing neighborhood-specific features. Thus, link prediction probabilities for pairs of nodes from the same region are likely to be similar. Such a property is motivated by the intuition that each node at a particular time belongs to some community, and nodes within the same community evolve similarly, but different communities might evolve differently or be in different stages of the same evolutionary pattern (e.g., one community is in the growth part of its life-cycle, while another is towards the tail end of the same life-cycle). Thus, the neighborhood around each node encapsulates this local community without requiring explicit community detection.

Pair-specific features. We assume that the pair-specific features $\mathrm{s}_t(i,j)$ come from a finite ordered set S; if not, they are discretized into such a set. For example, for any node $j \in \vec{N}_{t,p-1}(i)$ that was within two hops of i in the last $p-1$ timesteps, one may let $\mathrm{s}_t(i,j)$ record the number of common neighbors between i and j at time t, and the last time the $i \to j$ link appeared, up to a maximum of $p-1$ timesteps ago (lastlink); note that both are functions of $\vec{N}_{t,p-1}(i)$. Details of the features used in our experiments are provided later in Section 4.1.

Neighborhood-specific features. Let $d_t(i) = \{\eta_{i,t}(s), \eta^{+}_{i,t}(s); \forall s \in S\}$, where $\eta_{i,t}(s)$ are the number of node pairs in $N_{t-1}(i)$ with pair-specific feature vector s, and $\eta^{+}_{i,t}(s)$ the number of such pairs which were also linked by an edge in the next timestep t. In a nutshell, $d_t(i)$ tells us the chances of an edge being created in t given its features in $t-1$, averaged over the whole neighborhood $N_{t-1}(i)$—in other words, it captures the *change* of the neighborhood around i over one timestep. Note also that since $\eta_{i,t}(s)$ uses pair-specific feature vectors (which can look back $p-1$ timesteps) from $N_{t-1}(i)$, it is a function of $N_t(i)$, as is $d_t(i)$.

One can think of $d_t(i)$ as a contingency table indexed by the features s. Contingency tables are widely referred to as "datacubes" in the database community, and we will adopt this terminology, referring to $d_t(i)$ as a datacube, and a feature vector s as the "cell" s in the datacube with contents $(\eta_{i,t}(s), \eta^{+}_{i,t}(s))$. Finiteness of S is necessary to ensure that datacubes are finite-dimensional, which allows us to index them and quickly find nearest-neighbor datacubes.

2.2. *The estimator*

Our estimator of the function $g(\cdot)$ at time T is:

$$\tilde{g}_T(\psi_T(i,j)) = \frac{\sum\limits_{i',j',t'} \Gamma(\psi_T(i,j), \psi_{t'}(i',j')) \cdot Y_{t'+1}(i',j')}{\sum\limits_{i',j',t'} \Gamma(\psi_T(i,j), \psi_{t'}(i',j'))}, \tag{2.1}$$

$$\Gamma(\psi_T(i,j), \psi_{t'}(i',j')) = K(d_t\,(i)\,, d_{t'}\,(i')) \cdot \xi(\mathrm{s}_t\,(i,j)\,, \mathrm{s}_{t'}\,(i',j')). \tag{2.2}$$

Here, the kernel function $\Gamma(\psi_T(i,j), \psi_{t'}(i',j'))$ is further factored into a pair-specific part $\xi(\mathrm{s}_t\,(i,j)\,, \mathrm{s}_{t'}\,(i',j'))$ and a neighborhood-specific part $K(d_t\,(i)\,, d_{t'}\,(i'))$. We discuss these next.

Pair-specific factor. Let $\mathrm{dist}(s,s')$ denote the L_1 distance between features s and s', and let $n(s)$ denote the set of features at L_1 distance 1 from feature s. We define $\xi(\mathrm{s}_t\,(i,j)\,, \mathrm{s}_{t'}\,(i',j'))$ as

$$\begin{aligned} &\xi(\mathrm{s}_t\,(i,j)\,, \mathrm{s}_{t'}\,(i',j')) \\ &\quad := \frac{I\{\mathrm{s}_{t'}\,(i',j') = \mathrm{s}_t\,(i,j)\} + \zeta_T I\{\mathrm{dist}(\mathrm{s}_t\,(i,j)\,, \mathrm{s}_{t'}\,(i',j')) = 1\}}{1 + \zeta_T |n(\mathrm{s}_t\,(i,j))|}, \end{aligned} \tag{2.3}$$

where ζ_T is a bandwidth parameter which we will require to be $O(T^{-(1/2+\epsilon)})$ for some $\epsilon > 0$ in order to obtain consistency and distributional convergence. Note that $\xi(\mathrm{s}_t\,(i,j)\,, \mathrm{s}_{t'}\,(i',j')) \to I\{\mathrm{s}_{t'}\,(i',j') = \mathrm{s}_t\,(i,j)\}$ as $\zeta_T \to 0$. This factor can also be extended to features at L_1 distance two and so forth, while weighing those terms by powers of ζ_T.

Plugging in the definition of the kernel in Equation 2.1, we obtain the following interpretation of the estimator:

$$\tilde{g}_T(\psi_T(i,j)) = \frac{\sum_{i',t'} K\big(d_t(i), d_{t'}\big(i'\big)\big) \left(\eta^{+}_{i',t'+1}(\mathrm{S}_t(i,j)) + \zeta_T \sum_{s \in n(\mathrm{S}_t(i,j))} \eta^{+}_{i',t'+1}(s) \right)}{\sum_{i',t'} K(d_t(i), d_{t'}(i')) \left(\eta_{i',t'+1}(\mathrm{S}_t(i,j)) + \zeta_T \sum_{s \in n(\mathrm{S}_t(i,j))} \eta_{i',t'+1}(s) \right)}. \tag{2.4}$$

Useful intuition can be obtained by considering the case $\zeta_T = 0$. Here, given the query pair (i,j) at time t, we look inside cells for the query feature $s = \mathrm{s}_t\,(i,j)$ in all neighborhood datacubes, compute the average $\eta^{+}_{i',t'}\,(s)$ and $\eta_{i',t'}\,(s)$ in these cells after accounting for the similarities of the datacubes to the query neighborhood datacube, and use their quotient as the estimate of linkage probability. Letting $\zeta_T > 0$ provides an estimator that deals more effectively with sparsity by computing weighted averages of $\eta^{+}_{i',t'}\,(s)$ and $\eta_{i',t'}\,(s)$ over features s that are "close" to $\mathrm{s}_t\,(i,j)$.

Thus, the probability estimates are derived from historical instances where (a) the feature vector of the historical node pair matches the query, and (b) the local neighborhood is similar as well.

Neighborhood-specific factor. Now, we need a measure of the similarity $K(d_t\,(i)\,, d_{t'}\,(i'))$ between neighborhoods, with the goal of treating two neighborhoods as similar if they have similar probabilities of generating links between node pairs with feature vector s, for any $s \in S$. To this end we could simply compare point estimates $\eta^{+}_{\cdot}(s)/\eta_{\cdot}(s)$, but we also wish to account for the variance in these estimates. We achieve this by defining a similarity measure that has a Bayesian flavor:

$$\begin{aligned} K(d_t(i), d_{t'}(i')) &= e^{-D\left(d_t(i), d_{t'}(i')\right)/b_T} \quad (0 < b_T < 1) \qquad (2.5)\\ D(d_t(i), d_{t'}(i')) &= \sum_{s \in S} \mathrm{TV}(X, Y)\\ X &\sim \mathcal{B}\left(\eta^{+}_{i,t}(s), \eta_{i,t}(s) - \eta^{+}_{i,t}(s)\right)\\ Y &\sim \mathcal{B}\left(\eta^{+}_{i',t'}(s), \eta_{i',t'}(s) - \eta^{+}_{i',t'}(s)\right), \end{aligned}$$

where $\mathrm{TV}(X, Y)$ denotes the total variation distance between the distributions of X and Y, $\mathcal{B}$ is the beta distribution and $b_T \in (0, 1)$ is a bandwidth parameter. We will require $b_T = O(T^{-(1/2+\theta)})$ for some $\theta > 0$ to obtain appropriate rates when we study the consistency and distributional convergence of our estimator.

Thus, $K(d_t(i), d_{t'}(i'))$ is a discrete analog of a continuous kernel function (similar functions can be found in Aitchison and Aitken [2] and Wang and van Ryzin [33]). As is the case with continuous kernel functions, it has the property that as the bandwidth parameter $b_T \to 0$, it is equal to one if and only if $d_t(i) = d_{t'}(i')$, and zero otherwise.

Remarks. To better understand our choice of estimator, consider by way of contrast a simple estimator that computes the fraction of pairs for which the feature lastlink was equal to k at time t' and which formed an edge at time $t'+1$ (for $k = 1, 2, \ldots$). This approach suffers from two key problems that make it perform poorly on real-world graphs. First, it does not allow for local variations in the link-formation fractions, as would be expected for communities evolving differently within the same graph. We address this problem by maintaining a separate datacube for each local neighborhood. The second, more subtle, problem is the implicit assumption of stationarity—a node's link-formation probabilities are assumed to be time-invariant functions of the datacube features. This assumption does not allow for seasonal changes in linkage patterns, or for a transition from slow to fast growth, etc. Our model addresses this issue by finding historical neighborhoods from some previous time t' with datacubes similar to the query datacube, and uses their evolution from t' to $t'+1$ to predict link formation in the next time step for the current neighborhood. This helps us learn nonlinear trends.

Our estimator also has the virtue that it combats sparsity by aggregating data across similarly-evolving communities even if they are separated by graph distance and time. That said, sparsity remains a serious issue, and we provide a further discussion of sparsity in the following section.

Finally, note that we build the datacube so as to encode the recent change of a neighborhood, and not just the distribution of features in the neighborhood. Thus, for example, two neighborhoods may have the same datacube if the fraction of lastlink = 1 node pairs that formed an edge in the next timestep is the same in both neighborhoods, and not if they both merely had the same number of lastlink = 1 pairs. Thus, it is the change in link structure that drives the estimation of linkage probabilities. Moreover, two neighboring nodes may end up having very similar datacubes, and will end up forming links in a similar way, whereas very different datacubes will reflect the variations in link formation patterns among different communities.

2.3. Sparsity

If the graphs are sparse, or the time series short, the estimator may be unreliable in practice. There are two main reasons for this, which we discuss next.

First, a node i can have zero degree and hence an empty neighborhood. To cope with this issue, we consider the union of two-hop neighborhoods over the last $p-1$ timesteps (instead of just the current timestep) in constructing all pair-specific features. This reduces sparsity while retaining the property that the feature vector $\psi_t(i,j)$ is a function of the spatio-temporally bounded neighborhood $\vec{N}_{t,p}(i)$.

Second, and more problematically, the $\eta_{\cdot}(s)$ and $\eta_{\cdot}^{+}(s)$ values obtained from kernel regression can be small, yielding an estimated linkage probability $\eta_{\cdot}^{+}(s)/\eta_{\cdot}(s)$ that is unreliable numerically. We offer a threefold solution to this problem:

- *Using similar features:* The inner kernel ξ (Equation 2.3) combines $\eta_{\cdot}(s)$ and $\eta_{\cdot}^{+}(s)$ with a weighted average of the corresponding values for any s' that are "close" to s, the weights encoding the similarity between s' and s. Our estimator (Eq. 2.3) already does this for feature pairs within an L_1 distance of 1).
- *Accounting for uncertainty in ranking:* In determining a final ranking, instead of using $\eta_{\cdot}^{+}(s)/\eta_{\cdot}(s)$ directly, we use the lower end of the 95% Wilson score interval [34]. The node pairs that are ranked highest according to this "Wilson score" are those that have high estimated linkage probability $\eta_{\cdot}^{+}(s)/\eta_{\cdot}(s)$ *and* $\eta_{\cdot}(s)$ is high (implying a reliable estimate).
- *Smoothing via a prior:* We use a "backoff" smoothing procedure for the Wilson scores, in which the raw scores are smoothed against the scores obtained from a "prior" datacube, which is the average of all historical datacubes. The degree of smoothing depends on $\eta_{\cdot}(s)$. This can be thought of as a simple hierarchical model, where the lower level (set of individual datacubes) smooths its estimates using the higher level (the prior datacube).

3. Fast search using LSH

A naive implementation of the nonparametric estimator in Equation (2.4) computes kernel similarity between the query datacube and all n datacubes for each of the T timesteps for each prediction, which can be infeasibly slow for large graphs. To obtain a more computationally tractable estimator, we consider only the top-r closest neighborhoods (in terms of the largest kernel similarities). The value of r is a parameter of the algorithm; for our experiments we use $r=20$. What is needed to make this practical is a fast method (one that runs in sublinear time) to quickly find the top-r closest neighborhoods.

We achieve this by using locality sensitive hashing (LSH) [15]. Hashing is often used in databases for fast "table-lookups" or retrieving matching items from a large database. The key component is a hash function that maps a given "key" or object to a certain hash value. In order to search for a particular key, we compute the hash value and do a table lookup with this value. The concept of "locality sensitive" hashing refers to hash functions having the property that, with high probability, two "similar" data items are hashed to the same value. This facilitates approximate nearest neighbor search, and is suitable for high-dimensional spaces, where traditional nearest neighbor search techniques are often infeasible.

The standard LSH method operates on bit sequences, and maps sequences with small Hamming distance to the same hash bucket. In our setting, we must hash datacubes, and use the total variation distance metric. We make use of the fact that total variation distance between discrete distributions is half the L_1 distance between the corresponding probability mass functions. If we could approximate the probability distributions in each datacube cell with bit sequences, then the L_1 distance would just be the Hamming distance between these sequences, making our setting amenable to the use of standard LSH. We achieve this with three steps:

Conversion to bit sequence The key idea is to approximate the linkage probability distribution by discretization. We first discretize the range $[0, 1]$ (since we deal with probabilities) into B_1 buckets. For each bucket we compute the probability mass p falling inside it. This p is encoded using B_2 bits by setting the first $\lfloor pB_2 \rfloor$ bits to 1, and the others to 0. In this way the entire distribution (i.e., one cell) is represented by B_1B_2 bits. As a result the entire datacube can now be stored in $|S|B_1B_2$ bits. However, in all our experiments, datacubes were very sparse with only $M \ll |S|$ cells ever being non-empty (usually, 10–50); thus, we use only MB_1B_2 bits in practice. The Hamming distance between two pairs of MB_1B_2 bit vectors yields the total variation distance between datacubes (modulo a constant factor).

Distances via LSH We create a hash function by picking a uniformly random sample of k bits out of MB_1B_2. For each hash function, a hash table is created to store all datacubes whose hashes are identical in these k bits. We use ℓ such hash functions. A query datacube is first hashed using each of these ℓ functions. Then we create a *candidate set* containing $O(\max(\ell, \mathrm{r}))$ of distinct datacubes sharing any of these ℓ hashes. The total variation distance of these candidates to the query datacube is computed explicitly, yielding the closest matching historical datacubes.

Picking k The number of bits k is crucial in balancing accuracy versus query time: while a large k hashes all datacubes to their own hash bucket, returning a few or no matches to the query, a small k bunches many datacubes into the same bucket, decreasing the probability of finding the 'true' near neighbors. In the spirit of Indyk and Motwani [15], we do a binary search to find the k for which the average hash-bucket size over a query workload is just enough to provide the desired top-20 matches. We evaluate the accuracy of this approach in Section 4.

We conclude this section with two additional points. First, we never create the entire bit representation of MB_1B_2 bits explicitly; only the hashes need to be computed, taking $O(k\ell)$ time. Second, the main cost in the algorithm is in creating the hash table, which needs to be done once as a preprocessing step. Query processing is extremely fast and sublinear, since the candidate set is much smaller than the size of the training set.

4. Experiments

We start by introducing several baseline algorithms, and our evaluation metric. These baselines were picked carefully from previous work as being those that have yielded state-of-the-art performance in a range of link prediction tasks. In our first set of experiments we use simulated data to compare the performance of our algorithm to these baselines, focusing on situations involving seasonality in link formation. Second, we study the performance of our algorithm and the baselines on several real-world graphs: a sensor network, two co-authorship graphs, and a graph of Facebook employees. Finally, we investigate the computational scaling of our approach, comparing the improvement in runtime of the LSH-based algorithm to an exact algorithm, and investigating the effect of the LSH bit-size k on accuracy.

4.1. Experimental setup

We compare our nonparametric network inference algorithm (NNI) to the following baselines which, although quite naive, have proved difficult to beat in practice [18, 31]:

LL: ranks pairs using ascending order of *last time of linkage* [31].

CN (last timestep): ranks pairs using descending order of the number of *common neighbors* [18].

AA (last timestep): ranks pairs using descending order of the *Adamic-Adar* score [1], a weighted variant of common neighbors which it has been shown to outperform [18].

Katz (last timestep): extends CN to paths with length greater than two, but with longer paths getting exponentially smaller weights [16].

CN-all, AA-all, Katz-all: CN, AA, and Katz computed on *the union of all graphs until the last timestep.*

For NNI, we only predict on pairs which are in the neighborhood (generated by the union of two-hop neighborhoods of the last p timesteps) of each other. We deliberately used a simple feature set for NNI, setting $\mathrm{s}_t\,(i,j) = \{\mathrm{cn}_t(i,j), \ell\ell_t(i,j)\}$ (i.e., common neighbors and last-link) and not using any external “meta-data” (e.g., stock sectors, university affiliations, etc.). All feature values were binned logarithmically in order to combat sparsity in the tails of the feature distributions. Strictly speaking, our feature $\ell_t(i,j)$ should be capped at p. However, since the heuristic LL uses no such capping, for fairness, we used the uncapped "last time a link appeared" as the feature $\ell_t(i,j)$ for the pairs we predict on. The bandwidth b_T was picked by cross-validation.

For any graph sequence $(G_1, \ldots, G_T)$, we test link prediction accuracy on G_T for a subset $S_{>0}$ of nodes with non-zero degree in G_T. Each algorithm is provided training data up to and including timestep $T-1$, and must output, for each node $i \in S_{>0}$, a ranked list of nodes in descending order of probability of linking with i in G_T. For purposes of efficiency, we only require a ranking on the nodes that have ever been within two hops of i (call these the candidate pairs); all algorithms under consideration predict the absence of a link for nodes outside this subset. We compute the AUC score for predicted scores for all candidate pairs against their actual edges formed in G_T.

4.2. Simulations

In this section we compare NNI to the baseline algorithms using simulated data, focusing on seasonal patterns as an example of the kind of nonlinear behavior that may be difficult to capture with the heuristic methods. We simulated a model of Hoff et al [12] that posits an independently drawn "feature vector" for each node. Time moves over a repeating sequence of seasons, with a different set of features being "active" in each. Nodes with these features are more likely to be linked in that season, though noisy links also exist. The user features also change smoothly over time, to reflect changing user preferences.

Graph generation specifics. We generate feature vectors $u_{i,t} \in \mathbb{R}^6$ for node i at time t. Node pair $\{i,j\}$ has a link at time t if $u_{i,t}^T L_t u_{j,t} > 1$, where L_t is a matrix governing feature interactions. We define $u_{i,t}$ and L_t to model feature evolution and seasonal patterns, as follows.

Feature evolution: For every node i we generate a two feature vectors $a_i, b_i \sim \mathcal{N}(\mathbf{0}_6, I_{6\times 6})$, representing its features at the start $(t=1)$ and end $(t=T)$ of the simulation. At time $t \in \{1 \ldots T\}$, the node's features are given by $u_{i,t} = (c_t a_i + (1-c_t)b_i)/\sqrt{c_t^2 + (1-c_t)^2}$, where $c_t = \frac{T-t}{T-1}$. The normalization ensures identical variance of features at any timestep.

Seasonal patterns: We simulate a repeating cycle of three "seasons", with different features being important in different seasons. In particular, the six-dimensional feature vector is split into three groups of two features each, and in season $j = t \bmod 3$, we define

$$B_{k,\ell} = \begin{cases} \mu & \text{for } k,\ell \in \{2j+1, 2j+2\}, \\ 0 & \text{otherwise} \end{cases}$$

$$L_t = B + \sigma \frac{R + R^T}{2} \qquad \text{where } R \sim N(0,1)^{6\times 6},$$

where μ represents the signal and σ represents the noise. This feature interaction matrix L_t is then used to form links between node pairs, as mentioned above.

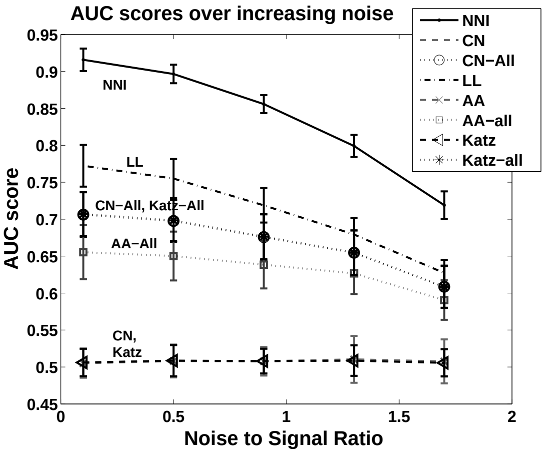


FIG 1. *Simulated graphs: Effect of noise.*

Results. We generated 100-node graphs over 20 timesteps using 3 seasons, and plotted AUC averaged over 10 random runs for several noise-to-signal ratios (Fig. 1). NNI consistently outperformed all other baselines by a large margin. Clearly, seasonal graphs have nonlinear linkage patterns: the best predictor of links at time T are the links at times $T-3$, $T-6$, etc., and NNI is able to learn this pattern. By contrast, CN, AA, and Katz are biased towards predicting links between pairs which are linked (or have short paths connecting them) at the previous timestep $T-1$; this implicit smoothness assumption makes them perform poorly; indeed, they behaved essentially as poorly as a random predictor (an AUC of 0.5).

Baselines LL, CN-all, AA-all and Katz-all use information from the union of all graphs until time $T-1$. Since the off-seasonal noise edges are not sufficiently large to form communities, most of the new edges come from communities of nodes created in season. This is why CN-all, AA-all and Katz-all outperform their "last-timestep" counterparts. As for LL, since links are more likely to come from the last seasons, it performed well, although poorly compared to NNI. Also note that the changing user features forces the community structures to change slowly over time; in our experiments, CN-all performed worse than it would were there was no change in the user features, since the communities stayed the same.

Table 1 summarizes the average AUC scores for graphs with seasonality, and also presents results for stationary data. In both cases, the noise was set to the smallest value in Fig. 1. For the stationary data, links formed in the last few timesteps of the training data are good predictors of future links, and so LL, CN, AA and Katz all performed very well. Interestingly, CN-all, AA-all and Katz-all were worse than their "last time-step" variants, presumably owing to the slow movement of the user features. As for NNI, it performed slightly better than all other methods for the stationary data, in addition to showing substantial improvements over the other methods for the seasonal networks.

TABLE 1
Average AUC for $T = 20$ timesteps

	Seasonal	Stationary
NNI	**$.91 \pm .01$**	**$0.99 \pm .005$**
LL	$.77 \pm .03$	$0.97 \pm .006$
CN	$.51 \pm .02$	$0.97 \pm .01$
AA	$.51 \pm .02$	$0.95 \pm .02$
Katz	$.50 \pm .02$	$0.97 \pm .01$
CN-all	$.71 \pm .03$	$0.86 \pm .03$
AA-all	$.65 \pm .04$	$0.71 \pm .04$
Katz-all	$.71 \pm .03$	$0.87 \pm .03$

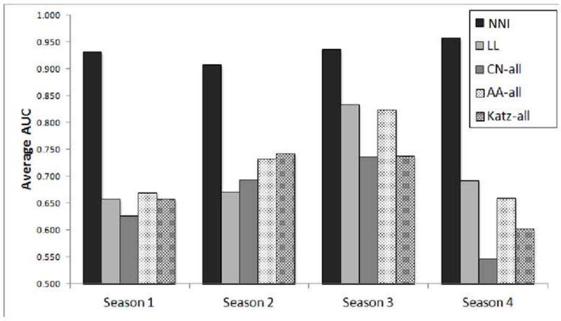


FIG 2. *AUC scores for a* ***periodic*** *sensor network.*

4.3. Real-world graphs

We begin by presenting results on a 24-node sensor network where each edge represents the successful transmission of a message[1]. We considered up to 82 consecutive measurements. These networks exhibit clear periodicity; in particular, a different set of sensors turn on and communicate during four different periods. Fig. 2 shows our results for these four periods averaged over several cycles. The maximum standard deviation, averaged over the periods, was .07. We do not show results for CN, AA and Katz, as they all performed no better than a random predictor. NNI significantly outperformed the baselines, confirming the results from the simulation experiments for seasonal graphs.

We also present results on three dynamic co-authorship graphs: the Physics "HepTh" community ($n = 14,737$ nodes, $e = 31,189$ total edges, and $T = 8$ timesteps), NIPS ($n = 2,865$, $e = 5,247$, $T = 9$), and authors of papers on Citeseer ($n = 20,912$, $e = 45,672$, $T = 11$) with "machine learning" in their abstracts. Each timestep considers 1–2 years of papers (so that the median degree at any timestep is at least 1). Finally we also considered a dynamic undirected network of Facebook employees over several weeks, where the nodes represent employees and edges are formed if one employee mentions another in a post. The network contains above five thousand nodes, and above 100,000 edges in total.

[1] http://www.select.cs.cmu.edu/data

Table 2

Average AUC for co-authorship and Facebook graphs

	NIPS	HepTh	Citeseer	Facebook
NNI	**.87**	**.89**	**.89**	.82
LL	.84	.87	**.90**	.81
CN	.74	.76	.69	.70
AA	.84	.87	**.90**	.71
Katz	.75	.83	.83	.78
CN-all	.56	.62	.70	**.87**
AA-all	.77	.83	.83	**.89**
Katz-all	.67	.71	.81	**.89**

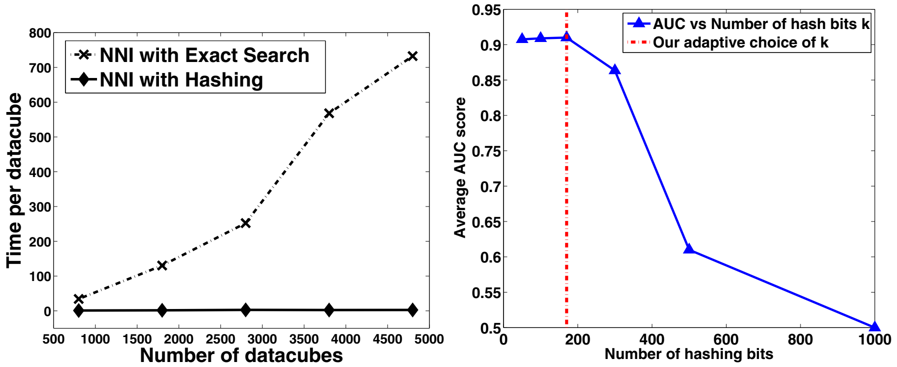
Table 2 shows the average AUC for all algorithms for the co-authorship graphs and the Facebook graph. For the co-authorship graphs, we do not expect to see seasonal variation, and we expect a relatively simple model to be effective; authors will tend to keep working with a similar set of co-authors over time. For such graphs, Tylenda et al. [31] have shown that LL is the best heuristic, and we replicate that result here. Our kernel-based approach, NNI, also performs well on these graphs, slightly outperforming LL. For the Facebook graph, employees in the same research group tend to post more messages mentioning each other, and hence algorithms working on all edges seen so far should intuitively pick up this community structure. This is indeed reflected in the AUC scores. CN-all, AA-all and Katz-all perform the best. These algorithms outperform NNI, primarily because they count paths through edges that exist in different timesteps, which is not allowed in our model.

In summary, for graphs having a seasonal trend, NNI is the best method by a large margin. For the co-authorship graphs, NNI remains the best algorithm, although LL is also effective. For the correlation graph, Katz-all is the best algorithm, but its performance is quite poor on the co-authorship graphs and the seasonal graphs. Overall, the performance of NNI dominates that of the other algorithms.

4.4. Model parameter selection

Here, we discuss our reasoning behind the selection of two model parameters: the size of the local neighborhood $N_t(i)$, and the temporal window length p.

Local neighborhood size: As mentioned in the beginning of Section 2, we define the local neighborhood of node i to be the set of nodes within two hops of i and all edges between them. Two is the minimum number of hops such that all common neighbors of the immediate neighbors of the central "ego" node i are included in the neighborhood of u. Since common neighbors is one of the best-performing baseline heuristics, the local neighborhood should subsume at least two hops. On the other hand, neighborhood sizes tend to grow exponentially with the number of hops [4], so significant computational resources would be required for more hops. Hence, we were led fairly decisively to the use of two hops.



(a) Time vs. #-datacubes

(b) AUC vs. hash bitsize k

Fig 3. *Time and accuracy using LSH.*

Length of temporal window: Recall that we model link formation using features derived from a node's local neighborhood over a time window of length p. Theoretically, for periodic data, the model learns the seasonal trend correctly as long as p is at least the period, since the local spatio-temporal neighborhood $N_{t,p}(i)(i)$ can then accurately identify the current "season". In order to verify this, we repeated the simulation experiment of Section 4.2 with $N = 100$ nodes, $\mu = 0.5$ (signal), $\sigma = 0.05$ (noise), a period of 3 ("seasons"), $T = 20$ timesteps, and varying the temporal window size $p \in \{2, \ldots, 7\}$. For $p = 2$, our model achieves an AUC score around 58% which is slightly better than random. For $p \geq 3$ (i.e., at least the periodicity of the data), the score is around 92%. This shows the robustness of the algorithm with respect the parameter p. In general one may wish to select this parameter via cross- validation, although the computational burden could be considerable.

4.5. Evaluation of LSH

We have found the use of LSH to be essential in our experimental work. In this section we provide quantitative support for this assertion.

Exact search vs. LSH. In Fig. 3(a) we plot the time taken to perform top-20 nearest neighbor search for a query datacube using simulated data. We fixed the number of nodes at 100, and increased the number of timesteps. As expected, the exact search time increases linearly with the total number of datacubes, whereas LSH searches in nearly constant time. Also, the AUC score of NNI with LSH is within 0.4% of that of the exact algorithm on average, implying minimal loss of accuracy from LSH.

In our experiments with real-world graphs, the query time per datacube using LSH was quite small: 0.3s for Citeseer, 0.4s for NIPS, 0.6s for HepTh, and 1.9s for Facebook. Exact search was infeasible for these large-scale graphs.

Number of Bits in Hashing. Fig. 3(b) shows the effectiveness of our adaptive scheme to select the number of hash bits (Section 3). For these experiments, we turned off the smoothing based on the prior datacube. As k increases, the accuracy goes down to 50%, as a result of the fact that NNI fails to find any matches of the query datacube. Our adaptive scheme finds $k \sim 170$, which yields the highest accuracy. Note also that larger k translates to fewer entries per hash bucket and hence faster searches, and thus our adaptive choice of k yields the fastest runtime performance as well.

5. Consistency of kernel estimator

We will now prove the consistency of $\tilde{g}$ (this section), and its asymptotic distribution (Sections 6 and 7) as $T \to \infty$. We note that the experiments in Section 4 demonstrated the accuracy of the estimator $\tilde{g}$ (Eq. (2.4)) in many settings, including for small T ($T = 20$ in our simulations); hence, asymptotics should be seen as providing a better understanding of our method rather than being a necessary condition for its applicability.

Recall that our model is:

$$Y_{t+1}(i,j)|\mathcal{G} \sim \text{Bernoulli}(g(\psi_t(i,j))), \tag{5.1}$$

where $\psi_T(i,j)$ equals $\{\mathrm{s}_t\left(i,j\right), d_t\left(i\right)\}$. Assume that all graphs have n nodes (n is finite). For a fixed node $q \in \{1,\ldots,n\}$, let Q represent the query datacube $d_T(q)$. We want to study the consistency of predictions for timestep $T+1$.

First, we provide some basic intuition for our results. The fact that edge probabilities depend on the feature vector $\psi_t(i,j)$, which itself is a function of the spatio-temporally bounded neighborhood $\vec{N}_{t,p}(i)$, means that the graph evolution process is Markovian; the graph at any timestep depends only on the previous $p+1$ graphs. This Markov Chain must eventually enter a closed communication class, and every state in that class will eventually be seen infinitely often. Thus, as $T \to \infty$, the graph at $T+1$ can be inferred from previous occurrences of the state at T, of which infinitely many instances will have been observed. The following proofs account for all the details: time spent in transient states, possible periodicity in the final communication class, proper choice of bandwidth parameters for the kernel, etc. Note that the $g(.)$ function is not learnt for all possible features $\psi_t(i,j)$, but only for those that appear in the communication class; however, this is enough to make predictions.

Rather than studying $\tilde{g}$ directly, it proves to be simpler to study a slightly different estimator which we show (in Lemma 5.1) to be asymptotically equivalent to $\tilde{g}$. Define $\widehat{g}_T(s,Q), \widehat{h}_T(s,Q)$ and $\widehat{f}_T(s,Q)$ as follows:

$$\widehat{g}_T(s,Q) = \frac{\widehat{h}_T(s,Q)}{\widehat{f}_T(s,Q)} \qquad (\text{where } s = \mathrm{s}_T\left(q,q'\right)) \tag{5.2}$$

$$\widehat{h}_T(s,Q) = \frac{1}{n(T-p)} \sum_{t=p}^{T-1} \sum_{i=1}^{n} K_{b_T}(d_t\left(i\right), Q)\eta^{+}_{i,t+1}\left(s\right)$$

$$\widehat{f}_T(s,Q) = \frac{1}{n(T-p)} \sum_{t=p}^{T-1} \sum_{i=1}^{n} K_{b_T}(d_t\left(i\right), Q)\eta_{i,t+1}\left(s\right).$$

Lemma 5.1. *Define $\tilde{g}_T(.)$ as in Equation 2.1, and $\widehat{g}_T(.)$ as in Equation 5.2. We have:*

$$|\tilde{g}_T(s,Q) - \widehat{g}_T(s,Q)| = O(\zeta_T)$$

Proof. Recall that $n(s)$ denotes the set of features at L_1 distance 1 from s. Let $k := |n(s)|$. We have:

$$\tilde{g}_T(s,Q) = \frac{\widehat{h}_T(s,Q) + C_T}{\widehat{f}_T(s,Q) + D_T},$$

where by virtue of the finiteness of number of features, η and η^+, we have:

$$C_T := \zeta_T \sum_{i,t} K_{b_T}(d_t\,(i)\,, Q) \sum_{s' \in n(s)} \eta^+_{it+1}(s') = O(\zeta_T).$$

Similarly, $D_T = O(\zeta_T)$. Also, note that both C_T and D_T are non-negative. Thus we have:

$$|\tilde{g}_T(s,Q) - \hat{g}_T(s,Q)| = \left| \frac{C_T \widehat{f}_T(s,Q) - D_T \widehat{h}_T(s,Q)}{(\widehat{f}_T(s,Q) + D_T)\widehat{f}_T(s,Q)} \right| = O(\zeta_T),$$

where the last step follows because both $\widehat{h}_T$ and $\widehat{f}_T$ are bounded and $\widehat{f}_T$ tends to some positive constant with probability tending to one as $T \to \infty$ (as shown in Theorem 5.2). □

The estimator $\widehat{g}_T$ is defined only when $\widehat{f}_T > 0$, which holds with probability tending to one as will be shown in the next theorem. The kernel was defined earlier as $K_{b_T}(d_t\,(i)\,, Q) = e^{-D(d_t(i),Q)/b_T}$, where the bandwidth b_T tends to 0 as $T \to \infty$, and $D(\cdot)$ is the distance function defined in Eq. (2.5). This has the following property:

$$\lim_{b_T \to 0} K_{b_T}(d_t\,(i)\,, Q) = \begin{cases} 1 & \text{if } d_t\,(i) = Q \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

From now on, we will drop the arguments s and Q and instead write g, $\widehat{g}_T$, $\widehat{f}_T$ and $\widehat{h}_T$ for simplicity. Our graph evolution model is Markovian; assuming each "state" to represent $p+1$ consecutive graphs, the next graph (and hence the next state) is a function only of the current state. The state space is also finite, since each graph has bounded size. Thus, the state space $\mathcal{S}$ may be partitioned into a set of transient states and $\bigcup_i C_i$, where C_i is an irreducible closed communication class, and there exists at least one C_i [9].

The Markov chain must eventually enter one of the (finitely many) communication classes. We will denote the time of entering some communication class by T_1, and the event by $\mathcal{E}_{T_1}$. We remind the reader that using simple arguments for finite state space Markov chains, it can be shown that the tail probability of T_1 decays geometrically (see [9]), leading to the finiteness of the first and second moments. Also let S_C denote the event $S_T \in C$, where S_t denotes the state of the Markov chain at time t. Thus $\mathcal{E}_{T_1} \cap S_C$ is the event that the chain enters class C at time T_1 and remains there henceforth.

Theorem 5.2 (Consistency)**.** *Let $b_T = o(1)$ as $T \to \infty$. For two fixed nodes $q, q' \in \{1, \ldots, n\}$, $\widehat{g}_T(s(q,q'), d_T(q))$ is well-defined with probability tending to one as $T \to \infty$. Also, $\widehat{g}_T(s(q,q'), d_T(q))$ is a consistent estimator of $g(s(q,q'), d_T(q))$, i.e., $\widehat{g}_T(s(q,q'), d_T(q)) \xrightarrow{P} g(s(q,q'), d_T(q))$ as $T \to \infty$.*

Proof. First, note that our query datacube is obtained at time T, and we are interested in the asymptotic behavior of the chain as $T \to \infty$. Since our Markov chain has a finite state space, the query datacube belongs to some closed communication class C with probability tending to one. Thus, as $T \to \infty$, the estimator's distribution is governed by that communication class. We prove our result in two parts; first we show that the convergence statement holds conditioned on S_C, for any communication class C; i.e., $P(|\widehat{g}_T - g| \geq \epsilon | S_C) \to 0$ as $T \to \infty$. Next, we have

$$P(|\widehat{g}_T - g| \geq \epsilon) \leq \sum_C P(|\widehat{g}_T - g| \geq \epsilon | S_C) P(S_C) + P(T_1 > T),$$

which implies $\limsup_{T\to\infty} P(|\widehat{g}_T - g| \geq \epsilon) = 0$, given the tail bound on T_1 and the fact that the first term is a sum over a finite number of terms, each converging to zero as $T \to \infty$. In what follows, we will give a proof of statistical consistency conditioned on S_C for any communication class C.

Define $B_T(s, Q, C) = E[\widehat{h}_T | S_C]/E[\widehat{f}_T | S_C] - g$. We have:

$$\widehat{g}_T - g = ([\widehat{h}_T - g\widehat{f}_T] - E[\widehat{h}_T - g\widehat{f}_T | S_C])/\widehat{f}_T + B_T E[\widehat{f}_T | S_C]/\widehat{f}_T. \tag{5.4}$$

Lemma 5.3 shows that $E[\widehat{f}_T | S_C] \to R_c$, R_c being a positive deterministic function of class C. Thus, B_T is asymptotically well defined. Also Lemma 5.9 shows that $\mathrm{var}(\widehat{f}_T | S_C)$ tends to 0 as $T \to \infty$. This along with Lemma 5.3 shows that, conditioned on S_C, $\widehat{f}_T \xrightarrow{P} R_c$, thus also proving that $\widehat{g}_T$ is asymptotically well defined for C.

Next, we will define the following:

$$\begin{aligned}
\widehat{h}_T(t) &:= \frac{1}{n} \sum_{i=1}^{n} K_{b_T}(d_t(i), Q) \eta^+_{i,t+1}(s), \\
\widehat{f}_T(t) &:= \frac{1}{n} \sum_{i=1}^{n} K_{b_T}(d_t(i), Q) \eta_{i,t+1}(s).
\end{aligned} \tag{5.5}$$

Note that $\widehat{h}_T$ and $\widehat{f}_T$ (Equation 5.2) equals $\sum_t \widehat{h}_T(t)/(T-p)$ and $\sum_t \widehat{f}_T(t)/(T-p)$ respectively. Also let

$$q_t := \widehat{h}_T(t) - E[\widehat{h}_T(t) | S_C] - g(\widehat{f}_T(t) - E[\widehat{f}_T(t) | S_C]). \tag{5.6}$$

Thus q_t is a bounded deterministic function of the state at time t. In Lemma 5.9 we prove that $\mathrm{var}(\sum_t q_t/\sqrt{T} | S_C) \to \sigma_c$ for some non-negative constant σ_c, as $T \to \infty$. Thus we have, $\mathrm{var}(\sum_t q_t/T | S_C) \to 0$, as $T \to \infty$. Since $E[q_t | S_C] = 0$, we have $\sum_t q_t/T \sim ([\widehat{h}_T - g\widehat{f}_T] - E[\widehat{h}_T - g\widehat{f}_T | S_C]) \xrightarrow{qm} 0$ conditioned on S_C.

Since convergence in quadratic mean implies convergence in probability, we have:

$$(\widehat{f}_T, [\widehat{h}_T - g\widehat{f}_T] - E[\widehat{h}_T - g\widehat{f}_T | S_C]) \overset{P}{\to} (R_c, 0) \quad \text{conditioned on} \quad S_C.$$

Using the continuous mapping theorem on $f(X,Y) = Y/X$ and the fact that $B_T = o(1)$ (Lemma 5.4) we have that, for any C such that $S_T \in C$, $\widehat{g}_T \overset{P}{\to} g$. □

The proof of the following lemma is deferred to the Appendix.

Lemma 5.3. *As $T \to \infty$, for some $R_c > 0$ (a deterministic function of class C),*

$$E[\widehat{f}_T(s,Q)|\mathcal{E}_{T_1}, S_C] \to R_c, \qquad\qquad E[\widehat{f}_T(s,Q)|S_C] \to R_c.$$

The following smoothness condition on g is introduced to ensure appropriate rates of convergence of the bias terms B_T.

Assumption 1. The function g satisfies the following smoothness condition with respect to the distance metric D: $|g(s, d_t(i)) - g(s, d_{t'}(j))| = O(D(d_t(i), d_{t'}(j)))$.

Lemma 5.4. *Define $B_T(s,Q,C) = (E[\widehat{h}_T(s,Q)|S_C] - gE[\widehat{f}_T(s,Q)|S_C])/E[\widehat{f}_T(s,Q)|S_C]$. If Assumption 1 holds, then we have $B_T = O(b_T)$. Since $b_T \to 0$ as $T \to \infty$, this implies $B_T = o(1)$.*

Proof Sketch. For $t \in [p, T-2]$, $i \in [1,N]$ and $s = \mathrm{s}_T(q,q')$, the numerator of B_T is an average of the terms:

$$A_t := E\left[K_{b_T}(d_t(i), Q)\eta^+_{i,t+1}(s)\,|S_C\right] - E\left[K_{b_T}(d_t(i), Q)\eta_{i,t+1}(s)\,|S_C\right] g(s,Q).$$

Using a further conditioning step on $\mathcal{E}_{T_1}$, we can show that the numerator of B_T can be upper bounded as:

$$\left|\sum_t A_t/T\right| \le \sum_t |E[K_{b_T}(d_t(i), Q)\eta_{i,t+1}(s)\,(g(s, d_t(i)) - g(s,Q))|S_C]|/T + o(1).$$

We now analyze each term in the average; i.e., terms of the form:

$$E\left[K_{b_T}(d_t(i), Q)\eta_{i,t+1}(s) \cdot (g(s, d_t(i)) - g(s,Q))\,|S_C\right].$$

This expectation is computed over all possible configurations of the neighborhoods $N_t(i)$ and $N_{t+1}(i)$. Since our neighborhood sizes are bounded (because n is bounded), the expectation is a sum over a finite number of terms.

We now use the smoothness assumption on g. Using $|g(s, d_t(i)) - g(s,Q)| = O(D(d_t(i), Q))$ and that $\eta_{i,t+1}(s)$ is finite for all T and Lemma 5.3, we have:

$$B_T = O\left(E[D(d_t(i), Q)e^{-D(d_t(i),Q)/b_T}|S_C]\right) = O(b_T),$$

which holds because for non-negative x, we have $xe^{-x/b_T} \le b_T/e$. □

We now show that the variance of $\widehat{f}_T$ and $\widehat{h}_T$ converge to zero. In order to upper bound the growth of variance terms, we make use of strong mixing. For a Markov chain S_t, define the strong mixing coefficients $\alpha(k) \doteq \sup_{|t-t'|\ge k}\{|P(A\cap$

$B) - P(A)P(B)| : A \in \mathcal{F}_{\leq t}, B \in \mathcal{F}_{\geq t'}\}$, where $\mathcal{F}_{\leq t}$ and $\mathcal{F}_{\geq t'}$ are the sigma algebras generated by events in $\bigcup_{i \leq t} S_i$ and $\bigcup_{i \geq t'} S_i$ respectively. Intuitively, small values of $\alpha(k)$ imply that states that are k apart in the Markov chain are almost independent. For bounded A and B, this also limits their covariance: $|\text{cov}(A, B)| \leq c\alpha(k)$ for some constant c [6]. Instead of proving that the variance of $\widehat{h}_T$ or $\widehat{f}_T$ converges to zero, we will prove that the variance divided by T converges to a non-negative constant. This is a stronger result that we will find useful in proving weak convergence in section 7.

We introduce some notation that will be used in stating the next few results. Let q_t denote a bounded deterministic function of the state of a finite state space Markov chain at time t. Also define $U_T := \sum_t q_t/\sqrt{T}$. Recall that our Markov chain will eventually hit one of the finitely many closed communication classes. Earlier we used S_C to define the event $\{S_T \in C\}$, by T_1 the time of entering some communication class, and the event by $\mathcal{E}_{T_1}$. We will denote the event of entering class C at time T_1 by $\mathcal{E}_{T_1} \cap S_C$. If C is aperiodic, then once inside C, the Markov chain gets arbitrarily close to the stationary distribution of C after some constant time M; we state this more formally in the following lemma, whose proof is deferred to the Appendix.

Lemma 5.5. *Consider an irreducible and aperiodic finite state Markov chain with probability transition matrix P, initial distribution π_0 and stationary distribution π. Let X_t be a random variable (with finite support) that is conditionally independent of all other states, given the state at time t. The expectation of X_t under the distribution at time t is denoted by $E[X_t|\pi_0]$. Let μ denote the expectation of X_∞ (i.e., the expectation with respect to π). There exists a constant $\lambda \in (0, 1)$, and a constant M such that, for all $t > M$, $\max_{x \in \mathcal{S}} \sum_{y \in \mathcal{S}} |P^t(x, y) - \pi(y)| = O(\lambda^t)$, and $|E[X_t|\pi_0] - \mu| = O(\lambda^t)$.*

Our estimators are weighted sums of $1, \dots, T$ variables; for $T_1 \leq T$, we will break this sum up into three parts, indexed by $1, \dots, T_1 - 1$, followed by $T_1, \dots, T_1 + M - 1$, and finally $T_1 + M, \dots, T$, where M is a constant. For $T_1 > T$, we will use the fact that T_1 has bounded first and second moments. Since we are interested in the behavior of the sum unconditionally, our analysis will consist of two steps of nested conditioning, the outer one obtained by conditioning on S_C, which in turn is obtained by analyzing the sum conditioned on $\mathcal{E}_{T_1} \cap S_C$. For ease of exposition we will assume C to be aperiodic. The more general case of cyclo-stationarity, which is similar in principle, is discussed in remark 5.10.

Lemma 5.6. $\text{var}(U_T|S_C) \to \sigma_c$ *as* $T \to \infty$, *for some constant* $\sigma_c \geq 0$.

Proof. We have $\text{var}(U_T|S_C) = E[\text{var}(U_T|\mathcal{E}_{T_1}, S_C)|S_C] + \text{var}(E[U_T|\mathcal{E}_{T_1}, S_C]|S_C)$. We prove that the first part converges to a non-negative constant σ_c (a deterministic function of C) (Lemma 5.7), and the second is asymptotically $o(1)$ (Lemma 5.8). □

Lemma 5.7. *For any finite integer k, we have*

$$\text{var}\left(\sum_{t \geq T_1 + M} q_t | \mathcal{E}_{T_1}, T_1 = k, S_C\right)/T \to \sigma_c \qquad \textit{for some } \sigma_c \geq 0 \tag{5.7}$$

$$\operatorname{var}\left(\sum_t q_t|\mathcal{E}_{T_1}, T_1 = k, S_C\right)/T \to \sigma_c \qquad \textit{for some } \sigma_c \geq 0. \tag{5.8}$$

For a Markov chain with a finite state space, we also have $E[\operatorname{var}(U_T|\mathcal{E}_{T_1}, S_C)|S_C] \to \sigma_c$ *for some* $\sigma_c \geq 0$.

Proof Sketch. For ease of exposition, for the proof sketch we assume there is only one communication class, which is aperiodic. Recall that T_1 is the time to hit the communication class. Once inside the communication class, irreducibility and aperiodicity implies geometric ergodicity (Lemma 5.5), which implies absolute regularity which in turn implies strong mixing with exponential decay [3]: $\alpha(k) \sim e^{-\beta k}$ for some $\beta > 0$. We can prove that for finite T_1, $\operatorname{var}(\sum_t q_t|\mathcal{E}_{T_1}, T_1 = k, S_C)/T = \operatorname{var}(\sum_{t\geq T_1+M} q_t|\mathcal{E}_{T_1}, T_1 = k, S_C)/T + o(1)$. So we focus on proving Equation 5.7. Denote $\sum_{t\geq T_1+M} q_t$ by P.

Recall that for our Markov chain, S_t involves $p+1$ graphs $(G_{t-p+1}, \ldots, G_{t+1})$. Since q_t is a function of S_t, it also depends on $p+1$ graphs. Hence, the distance $\operatorname{dist}(t,t')$ between two sigma-algebras $\mathcal{F}_{\leq t}$ and $\mathcal{F}_{>t'}$ is defined as $\max(t'-t-(p+1), 0)$. Now we can write $\operatorname{var}(P|\mathcal{E}_{T_1}, S_C)$ as

$$\operatorname{var}(P|\mathcal{E}_{T_1}, S_C) = 2\sum_{t\geq T_1+M}\ \sum_{\operatorname{dist}(t,t')=0}^{T-t} \operatorname{cov}(q_t, q_{t'}|\mathcal{E}_{T_1}, S_C).$$

Since the number of states at distance 0 is $O(p+1)$, and at distance ≥ 1 is $O(1)$, for constants $\{c_k, k \geq 0\}$ we have,

$$\sum_{\operatorname{dist}(t,t')=0}^{T-t} |\operatorname{cov}(q_t, q_{t'}|\mathcal{E}_{T_1}, S_C)| \leq \sum_{k=0}^{\infty} c_k\alpha(k) = O\left(\sum_k e^{-\beta k}\right) = O(1).$$

This shows that the above sum converges to some constant a_t. Since $t \geq T_1+M$, the chain will get arbitrarily close to stationarity, and $a_t \to \sigma_c$ for some constant σ_c. Thus $\operatorname{var}(P|\mathcal{E}_{T_1}, S_C)/T$ is asymptotically equivalent to $\sum_t a_t/T$, which also converges to σ_c as $T \to \infty$. Since, for all T, $\operatorname{var}(P|\mathcal{E}_{T_1}, S_C)/T$ is non-negative, σ_c is also non-negative. This proves Equation 5.7. Thus Equation 5.8 is proved, and also, since T_1 has finite first and second moments for a finite state space Markov chain, $E[\operatorname{var}(\sum_t q_t|\mathcal{E}_{T_1}, S_C)|S_C]/T$ converges to σ_c, as $T \to \infty$. □

It remains to analyze $\operatorname{var}(E[U_T|\mathcal{E}_{T_1}, S_C]|S_C)$ in the variance decomposition. Using Lemma 5.5 we can prove that $|E[U_T - \mu_c|S_C, \mathcal{E}_{T_1}]|$ approaches zero at a geometric rate as $T \to \infty$, where μ_c denotes the expectation of q_t under the stationary distribution in communication class C. This implies the following lemma, which is proved in the Appendix.

Lemma 5.8. $\operatorname{var}(E[U_T|\mathcal{E}_{T_1}, S_C]|S_C) = o(1)$.

Lemma 5.9. $\operatorname{var}(\widehat{h}_T|S_C)$ *and* $\operatorname{var}(\widehat{f}_T|S_C)$ *tend to* 0 *as* $T \to \infty$.

Proof. The result follows by applying Lemma 5.6 with $q_t(.)$ equal to $\sum_i K_{b_T}(d_t(i), Q)\eta^+_{i,t+1}(s)/n$ and $\sum_i K_{b_T}(d_t(i), Q)\eta_{i,t+1}(s)/n$ respectively. □

Remark 5.10. Recall that Lemma 5.7 was obtained under the assumption that C is aperiodic. The case of periodic C implies cyclo-stationarity; i.e., the chain S_{t+kd} approaches stationarity as $k \to \infty$. Hence, for periodic C (with period d) we consider $\mathcal{M}'$, which is a Markov chain where each transition corresponds to d transitions of the original chain. Now, $\mathcal{M}'$ is irreducible and aperiodic (since C was irreducible and had period d). A state S'_t in $\mathcal{M}'$ started at S_1 simply corresponds to the old state S_{td+1} in $\mathcal{M}$. Now, $1/\sqrt{T}\sum_{t=1}^{T} q_t$ can be written as $1/\sqrt{T}\sum_{t=1}^{\lfloor T/d \rfloor} q'_t + o_P(1)$, where $q'_i := \sum_{j=id+1}^{(i+1)d} q_j$ is the sum of d consecutive random variables. Since, q'_t is independent of all other q's conditioned on S'_t, S'_{t+1}, we have:

$$\begin{aligned}
\operatorname{cov}(q'_t, q'_{t+k}) &= E[E[q'_t q'_{t+k}|S'_{t+1}, S'_{t+k}]] - E[q'_t]E[q'_{t+k}] \\
&= E[E[q'_t|S'_{t+1}]E[q'_{t+k}|S'_{t+k}]] - E[E[q'_t|S'_{t+1}]]E[E[q'_{t+k}|S'_{t+k}]] \\
&= \operatorname{cov}(E[q'_t|S'_{t+1}], E[q'_{t+k}|S'_{t+k}]) = O(\alpha(k-1)).
\end{aligned} \tag{5.9}$$

The last step uses the fact that the q'_t are bounded. Now, $E[\operatorname{var}(1/\sqrt{T}\sum_{t=1}^{\lfloor T/d \rfloor} q'_t|\mathcal{E}_{T_1}, S_C)]$ can again be shown to converge to some non-negative constant using a slight modification of the argument in Lemma 5.7. The $o_P(1)$ remainder of $1/\sqrt{T}\sum_{t=1}^{T} q_t$ can be shown to be negligible via a simple application of the Cauchy-Schwartz inequality. A detailed proof of Lemma 5.7 using this idea can be found in the Appendix.

As for $E[U_T|\mathcal{E}_{T_1}, S_C]$ in the cyclic case, we simply have to apply Lemma 5.5 for each of the d cyclic classes. For the i^{th} cyclic class, q_{T_1+kd+i} is independent of all states (in that cyclic class) given S_{T_1+kd+i}. Hence there exists M_i, and $\lambda_i \in (0,1)$, such that for all k with $kd+i > M_i$, $|E[q_{T_1+kd+i}|\mathcal{E}_{T_1}, S_C] - \mu_i| = O(\lambda_i^k)$, thus proving Lemma 5.5 for a periodic C. This again proves Lemma 5.8 for the case where C is periodic.

6. Stein's method for graphical data

Our estimators, and indeed many kernel estimators, involve weighted sums of dependent variables. While their distributional convergence can be studied using existing results on ergodic Markov chains, we take a different approach, based on an adaptation of Stein's method to the setting of graphs.

We begin with a brief introduction to Stein's method. The method reposes on the following key lemma [5], which provides a characterization of the normal distribution:

Lemma 6.1 (Stein's Lemma)**.** *If W has a standard normal distribution, then*

$$Ef'(W) = E[Wf(W)], \tag{6.1}$$

for all absolutely continuous functions $f : \mathbb{R} \to \mathbb{R}$ with $E|f'(Z)| < \infty$. Conversely, if Equation 6.1 holds for all bounded, continuous and piecewise continuously differentiable functions f with $E|f'(Z)| < \infty$, then W has a standard normal distribution.

Recall that the Wasserstein distance between a mean zero, unit variance random variable W and a standard normal variate Z is defined as $\sup_{h\in\mathcal{H}} |Eh(X) - Eh(Z)|$, where $\mathcal{H} := \{h : |h(x) - h(y)| \leq |x - y|\}$. Weak convergence of W to Z can be established by showing that the Wasserstein distance converges to zero. Now, Stein's Lemma (6.1) shows that $W \overset{d}{=} Z$ if $|Ef'(W) - E[Wf(W)]|$ equals zero for appropriate choices of f. This key observation leads to the Stein Equation:

$$f'(W) - Wf(W) = h(W) - E[h(Z)]. \tag{6.2}$$

It can be shown that the solution to the Stein Equation, for $h \in \mathcal{H}$, satisfies $\|f\| \leq 2$, $\|f'\| \leq 2$, $\|f''\| \leq \sqrt{2/\pi}$ [5]. Thus, instead of dealing with $E[h(W)] - E[h(Z)]$ we need to show that $|E[f'(W) - Wf(W)]|$ is small (where f satisfies the aforementioned conditions); this is an easier quantity to analyze.

The existing application of Stein's method to sums of weakly dependent random variables has focused on marginal-independence structures that can be captured by a bounded-degree dependency graph [26]. In this section, we relax the requirement of marginal independence by allowing arbitrary dependency structures among the summed variables as long as certain conditions on strong mixing coefficients $\alpha(k)$ hold. (See also Sunklodas [30] for a similar approach to ours for chain-structured dependencies; he obtains a slightly tighter bound than ours at the expense of a more complex proof.)

Our approach proceeds by bounding the Wasserstein distance between the (appropriately scaled and centered) sum W of the dependent variables and a standard normal variate Z in terms of $\alpha(k)$ and the degree of dependence of the random variables. We then show that this bound tends to zero for our estimators, demonstrating convergence to a normal distribution and yielding a rate of convergence as a by-product. We note that although we use this to prove normal convergence for a cyclo-stationary Markov chain, it can potentially be used for more general dependence structures, as long as suitable strong mixing properties are available.

We let T denote the total number of variables in our model. Let $Y_i, \{i = 1, \ldots T\}$ be bounded, ($|Y_i| \leq B$), mean-zero random variables. Let ${\sigma_T}^2$ denote the variance of $\sum_i^T Y_i$; assume $0 < \sigma_T < \infty$ for all T. Define $X_i = Y_i/\sigma_T$, where $|X_i| \leq B/\sigma_T$. Let $W := \sum_i^T X_i$, and $\gamma_T = T/\sigma_T$. We will assume that the index set underlying the random variables $\{X_i\}$ is endowed with a distance metric, $\text{dist}(i, j)$. This can be the geodesic distance if the variables are connected via a graph structure or the absolute difference in time indices in a time series model, etc. Let $N_m(i)$ denote the set of nodes at distance m from node i; similarly let $N_{\leq k}(i)$ and $N_{>k}(i)$ respectively denote the set of nodes within distance k and at a distance greater than k from node i. Now, let $|N_{\leq k}|$ denote $\max_i |N_{\leq k}(i)|$.

We need a notion of strong mixing in a network setting. Define the strong mixing coefficients $\alpha(k) \doteq \sup_{X_i, X_j}\{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}(X_i), B \in \mathcal{F}(X_j), \text{dist}(i, j) \geq k\}$, where $\mathcal{F}(X)$ is the sigma algebra generated by the random variable X. A similar proposal for strong mixing in random fields can be found in [22]. Let τ_k denote the tail sum $\sum_{m>k} |N_m|\alpha(m)$. We are now ready to state the main result.

Lemma 6.2. *The Wasserstein distance $d_W(W,Z)$ between W and the standard normal random variable Z is upper bounded as follows:*

$$d_W(W,Z) \le \min_{k\le T}\left(c_1 B^3 \gamma_T \left(\frac{|N_{\le k}|}{\sigma_T}\right)^2 + c_2 B \gamma_T \alpha(k) \right. \tag{6.3}$$
$$\left. + B^2\sqrt{c_3\left(\frac{\gamma_T \tau_k}{\sigma_T}\right)^2 + c_4\gamma_T\left(\frac{|N_{\le k}|}{\sigma_T}\right)^3 + c_5\gamma_T\frac{\tau_k}{\sigma_T}\left(\frac{|N_{\le k}|}{\gamma_T}\right)^2}\right),$$

where c_1, c_2, c_3, c_4, c_5 are constants.

Proof sketch. We will give a brief proof sketch here, and provide the full proof in the Appendix. We want to bound $|E[f'(W) - Wf(W)]|$. We shall repeatedly break up W into two parts: $W_i = \sum_{j\in N_{>k}(i)} X_j$ being the contribution from all nodes with distance more than k from some node i, and the remainder from nodes "close to" i. In classical analysis of dependency graphs, X_i and W_i are independent; in contrast, in our case we only have $\text{cov}(X_i, W_i) = O(\alpha(k))$. Here, k is a parameter that shall be picked later to optimize the bound. Since $W = \sum_{i=1}^T X_i$,

$$|E[f'(W) - Wf(W)]| \le \underbrace{\left|E\left[f'(W)\left(1 + \sum_i X_i(W_i - W)\right)\right]\right|}_{(A1)}$$
$$+ \underbrace{\left|E\left[\sum_i X_i(W_i - W)f'(W) + \sum_i X_i f(W)\right]\right|}_{(A2)}.$$

Using Taylor expansion the term $(A2)$ can be further bounded by

$$(A2) \le \frac{\|f''\|}{2} E\left|\sum_i X_i(W_i - W)^2\right| + \left|E\left[\sum_i X_i f(W_i)\right]\right|.$$

The first term of this result again can be bounded using the AM-GM inequality by $c_1 B^3 \frac{T|N_{\le k}|^2}{\sigma_T{}^3}$, where c_1 is a constant. Recall that $|N_{\le k}|$ upper bounds the size of the neighborhood of k hops. The second part of (A2) now is bounded by $c_2 B \frac{T\alpha(k)}{\sigma_T}$, using the usual relationship between covariances and strong mixing coefficients. Thus the overall bound on $(A2)$ is as follows:

$$(A2) \le c_1 B^3 \frac{T|N_{\le k}|^2}{\sigma_T{}^3} + c_2 B \frac{T\alpha(k)}{\sigma_T}.$$

Now we need to bound $(A1)$. Denote $P_T = \sum_i X_i(W_i - W)$. Note that *if X_i and W_i were independent*, we would have $E[P_T] = -E(W^2) = -1$, since W is centered and scaled appropriately. For us however $E[P_T]$ does not equal -1; instead it becomes smaller as we increase k. We thus bound $(A1)$ as follows:

$$(A1) \le \|f'\|\sqrt{E[1+P_T]^2} \le \|f'\|\sqrt{(1+E[P_T])^2 + \text{var}(P_T)}.$$

Now note that $|1 + E[P_T]| = |E[\sum_i X_i W_i]|$. Since $E[X_i] = 0$,

$$\left|E\left[\sum_i X_i W_i\right]\right| = \left|\sum_i \sum_{j \in N_{>k}(i)} \mathrm{cov}(X_i, X_j)\right| \leq c'' B^2/{\sigma_T}^2 \sum_i \sum_{m>k} \alpha(m)|N_m|,$$

using the fact that $N_{>k}(i) = \bigcup_{m>k} N_m(i)$, and for all $j \in N_m(i)$, $\mathrm{cov}(X_i, X_j) = O(\alpha(m)/{\sigma_T}^2)$. We upper bound $|1 + E[P_T]|$ by $c'' B^2 T \tau_k/{\sigma_T}^2$. Using similar arguments (see Appendix) we upper bound $\mathrm{var}(P_T)$ by $8B^4 T|N_{\leq k}|^3/{\sigma_T}^4 + 16B^4 T|N_{\leq k}|^2 \tau_k/{\sigma_T}^4$.

Putting the pieces together and using $\gamma_T = T/\sigma_T$ we see that

$$\begin{aligned} d_W(W, Z) &\leq (A2) + \|f'\|\sqrt{(1 + E[P_T])^2 + \mathrm{var}(P_T)} \\ &\leq c_1 B^3 \gamma_T \left(\frac{|N_{\leq k}|}{\sigma_T}\right)^2 + c_2 B \gamma_T \alpha(k) \\ &\quad + B^2 \sqrt{c_3 \left(\frac{\gamma_T \tau_k}{\sigma_T}\right)^2 + c_4 \gamma_T \left(\frac{|N_{\leq k}|}{\sigma_T}\right)^3 + c_5 \gamma_T \frac{\tau_k}{\sigma_T}\left(\frac{|N_{\leq k}|}{\sigma_T}\right)^2}. \end{aligned}$$

The result is obtained by optimizing the upper bound over $k \leq T$. □

Next, we present a sufficient condition for the Wasserstein distance to vanish asymptotically, implying convergence of W to a standard normal.

Lemma 6.3. *$W \to \mathcal{N}(0, 1)$ as $T \to \infty$ if the following conditions hold:*

1. *$\gamma_T \to \infty$.*
2. *There exists a sequence $k(T) \to \infty$ such that the following are satisfied:*
   (a) $\gamma_T \alpha(k(T)) \to 0$
   (b) $\gamma_T \frac{\tau_{k(T)}}{\sigma_T} \to 0$
   (c) $\gamma_T (\frac{|N_{\leq k(T)}|}{\sigma_T})^2 \to 0$.

Proof. The above conditions imply that $\alpha(k(T)) \to 0$, $\frac{\tau_{k(T)}}{\sigma_T} \to 0$, and $(\frac{|N_{\leq k(T)}|}{\sigma_T})^2 \to 0$ (and thus $\frac{|N_{\leq k(T)}|}{\sigma_T} \to 0$ as well) as $T \to \infty$. Hence the product of two vanishing sequences, $\gamma_T (\frac{|N_{\leq k(T)}|}{\sigma_T})^2 \times \frac{|N_{\leq k(T)}|}{\sigma_T}$, also vanishes. Similarly $(\gamma_T \frac{\tau_k}{\sigma_T})(\frac{|N_{\leq k(T)}|}{\sigma_T})^2$ also vanishes as $T \to \infty$. Thus, all terms on the right hand side of Eq. (6.3) vanish, thus proving $W \xrightarrow{d} \mathcal{N}(0, 1)$. □

7. Weak convergence of our estimator

In this section we bring together the results from the previous two sections to establish weak convergence of our estimator.

Recall that our estimator $\tilde{g}_T$ is defined in Equation 2.4. Recall also the definitions of $\widehat{h}_T(t)$, $\widehat{f}_T(t)$ and q_t from Equations 5.5 and 5.6. From Lemma 5.1 we have $|\sqrt{T}(\tilde{g}_T - \widehat{g}_T)| = O(\sqrt{T}\zeta_T)$, where ζ_T denotes the bandwidth for the pair-specific kernel function (see Equation 2.3). Hence, with $\zeta_T = T^{-(1/2+\epsilon)}$ for some $\epsilon > 0$, we see that $\sqrt{T}(\tilde{g}_T - \widehat{g}_T) \overset{a.s.}{\to} 0$. We will show (in Proposition 7.1) that under suitable conditions $\sqrt{T}(\widehat{g}_T - g)$ converges to a mean-zero normal distribution. Hence, we also have the same normal distribution as the limit of $\sqrt{T}(\tilde{g}_T - g)$ under the same conditions.

Proposition 7.1. *Let Assumption 1 hold. If $\sigma_c > 0$, and $b_T = T^{-(1/2+\theta)}$ for some $\theta > 0$, then:*

$$\textit{Conditioned on } \; S_C, \qquad \sqrt{T}(\widehat{g}_T - g) \overset{d}{\to} \mathcal{N}(0, \sigma_c^2/R_c^2) \qquad \textit{As } T \to \infty.$$

where S_T is the state of the Markov chain at time T.

Proof. From Equation 5.4 we see that $\sqrt{T}(\widehat{g}_T - g)$ equals $(\sum_t q_t/\sqrt{T})/\widehat{f}_T + (E[\widehat{f}_T|S_C]/\widehat{f}_T)/\sqrt{T}B_T$. Using the following lemma (Lemma 7.2) we know that the numerator of the first term converges to a $\mathcal{N}(0, \sigma_c^2)$ distribution. Using Lemmas 5.9 and 5.3 we have $\widehat{f}_T \overset{P}{\to} R_c$ for a positive constant R_c, conditioned on S_C. Hence using Slutsky's lemma the first part converges conditionally to $\mathcal{N}(0, \sigma_c^2/R_c^2)$. Also, $E[\widehat{f}_T|S_C]/\widehat{f}_T$ converges to one in probability conditioned on S_C. Finally, invoking Lemma 5.3 and Lemma 5.4 we see that since $B_T = O(b_T)$, for $b_T \sim T^{-(1/2+\theta)}$, the second part is $o_P(1)$. Now, Slutsky's lemma and the continuous mapping theorem yield the statement of the proposition. □

Lemma 7.2. *Under Assumption 1 and assuming $\sigma_c > 0$,*

$$\textit{Conditioned on } \; S_C, \qquad \sum_t q_t/\sqrt{T} \overset{d}{\to} \mathcal{N}(0, \sigma_c^2) \qquad \textit{As } T \to \infty.$$

Proof Sketch. We prove this in two steps. First we show that conditioned on $\mathcal{E}_{T_1} \cap S_C$, a related quantity $\sum_{t \geq T_1+M} p_t/\sqrt{T}$ converges to $\mathcal{N}(0, \sigma_c^2)$ (Lemma 7.3). This along with a geometric bound on the tail probability of T_1 for finite state space Markov chains concludes the proof. We defer the details to the Appendix. □

Lemma 7.3. *Define $p_t := [\widehat{h}_T(t) - g\widehat{f}_T(t)] - E[\widehat{h}_T(t) - g\widehat{f}_T(t)|\mathcal{E}_{T_1}, S_C]$. Under Assumption 1 and assuming $\sigma_c > 0$, for any finite T_1, we have:*

$$\sum_{t \geq T_1+M} p_t/\sqrt{T} \overset{d}{\to} \mathcal{N}(0, \sigma_c^2) \qquad \textit{conditioned on } \mathcal{E}_{T_1} \cap S_C \textit{ as } T \to \infty.$$

Proof Sketch. First we prove that, for a sequence $k(T) = c \log T$ for a properly chosen c, the conditions in Lemma 6.3 are satisfied for

$$W_T := \left(\sum_{t \geq T_1+M} p_t \right) \Bigg/ \sqrt{\operatorname{var}\left(\sum_{t \geq T_1+M} p_t | \mathcal{E}_{T_1}, S_C \right)}.$$

We also show that for this value of k, the upper bound on the Wasserstein distance in Lemma 6.2 is $O(\log^2(T)/T)$. Now Lemma 6.1 gives:

$$W_T \overset{d}{\to} \mathcal{N}(0,1) \qquad \text{conditioned on } \mathcal{E}_{T_1} \cap S_C.$$

However, for finite values of T_1, $\mathrm{var}(\sum_{t \geq T_1+M} p_t | \mathcal{E}_{T_1}, S_C)/T \to \sigma_c^2$ (Lemma 5.7 and Equation 5.7). Thus, the additional assumption of $\sigma_c > 0$ proves the result. The details are deferred to the Appendix. □

Remark 7.4. Proposition 7.1 shows that, under some weak assumptions, W_T converges to a standard normal distribution conditioned on S_C. Since there are a finite number of closed communication classes, unconditionally W_T converges to a mixture of zero-mean Gaussians, the mixture proportions being determined by the probability of reaching the communication classes from the start state.

Remark 7.5. We have established weak convergence for the case where C is aperiodic. However, as in Remark 5.10, we can consider $\mathcal{M}'$, which is a Markov chain where each transition corresponds to d transitions of the original chain. Again, any sum of the form $\sum_{t=1}^{T} q_t/\sqrt{T}$ can be written as $1/\sqrt{d}(\sum_{t=1}^{\lfloor T/d \rfloor} q'_t/\sqrt{T/d} + o_P(1))$. q'_t now denotes the sum of the d consecutive q_t's. For $q'_i := \sum_{j=id+1}^{(i+1)d} q_j$, we have $\mathrm{cov}(q'_t, q'_{t+k}) = O(\alpha(k-1))$ using Equation 5.9. Thus the first sum again brings us to the irreducible aperiodic setting (with a slightly modified distance function), and hence normal convergence can be established.

8. Related work

Existing work on link prediction in dynamic networks can be broadly divided into two categories: link prediction based on generative models and link prediction based on structural features.

A substantial amount of work has gone into the development of generative models of graph structure based on the formalism of Markov random fields, loglinear models or other graphical models [8, 10, 17, 28, 13, 29, 32]. For example, Hanneke and Xing [10] present a dynamic loglinear model based on evolution statistics such as "edge stability," "reciprocity" and "transitivity." Fu et al. [8] propose an extension of the mixed membership block model to allow a linear Gaussian trend in the model parameters. Zhou et al. [35] present a nonparametric approach to estimating a time-varying Gaussian graphical model where the covariance matrix changes smoothly over time. The discrete analog of this is considered in [17], where the goal is to learn the latent structures of evolving graphs from a time series of node attributes. The static model of Raftery et al. [24] is extended by Sarkar and Moore [28] by allowing smooth transitions in latent space. All of these models have the virtue of a clean probabilistic formulation such that link prediction can be cast in terms of Bayesian posterior inference. Obtaining this posterior is, however, often infeasible in large-scale graphs. Moreover, these models often make strong model assumptions, not only for the graph structure but also for the network dynamics, which is often modeled as linear.

Alternatives to generative models generally revolve around the definition of various static features that aim to capture structural properties of graphs. These are extended to the dynamic setting via heuristics or via autoregressive modeling. For example, Huang and Lin [14] propose a linear autoregressive model for link prediction and investigate simple combinations of static graph-based similarity measures (e.g., Katz, common neighbors) with their autoregressive model to capture transitive similarities in networks. A similar parametric approach can be found in Richard et al. [25], where a vector autoregressive model was used for link prediction in dynamic graphs. The authors assume a low rank structure of the graph adjacency matrices and propose proximal methods for inference.

Tylenda et al. [31] examine simple temporal extensions of existing static measures. As we have noted earlier, these methods have the virtue of being applicable to large-scale graphs. They also tend to yield surprisingly good performance. Our work falls into this general category, while going beyond existing work by providing a formal statistical treatment of link prediction as a nonparametric estimation problem.

We conclude this section with a brief discussion on relevant research on nonparametric bootstrap estimators in strong mixing random fields and Markov processes. While these works are not relevant to the link prediction aspect of our work, they are similar because the estimation uses local resampling methods thereby retaining the dependency structure of the data. In the context of strong mixing random fields Politis and Romano [23] consider a blocks of blocks resampling method for estimating asymptotically accurate confidence intervals for parameters of the joint distribution of the random field. Nonparametric bootstrap algorithms have also been applied successfully to the area of computer vision. E. and J. [7] show that one such heuristic algorithm for texture synthesis can be formally framed as a resampling technique for stationary random fields, and prove consistency properties of it under broad conditions. In the context of stochastic processes with an autoregressive structure, Paparoditis and Dimitris [20] present the "local bootstrap" algorithm, which implicitly estimates the distribution of the one-step transition in the underlying Markov process and generates the bootstrap replicates using this estimated distribution.

9. Conclusions

In this paper we proposed a nonparametric model (NNI) for link prediction in dynamic networks, and showed that it performs as well as the state of the art for several real-world graphs, and exhibits important advantages over them in the presence of nonlinearities such as seasonality patterns. NNI also allows us to incorporate features external to graph topology into the link prediction algorithm, and its asymptotic convergence to the true link probability is guaranteed under our fairly general model assumptions. In addition, we show how to make NNI computationally tractable via the use of locality sensitive hashing. Together, these make NNI a useful tool for link prediction in dynamic networks.

Appendix

A.1. Statement and proofs of results from Section 5

Lemma 5.3. *As $T \to \infty$, for some $R_c > 0$ (a deterministic function of class C),*

$$E[\widehat{f}_T(s,Q)|\mathcal{E}_{T_1}, S_C] \to R_c, \qquad E[\widehat{f}_T(s,Q)|S_C] \to R_c.$$

Proof. Let ϵ denote the minimum distance between two datacubes that are not identical; since the set of all possible datacubes is finite, $\epsilon > 0$. $E[\widehat{f}_T(s,Q)|\mathcal{E}_{T_1}, S_C]$ is an average of terms $E[K_{b_T}(d_t\,(i)\,,Q)\eta_{i,t+1}\,(s)\,|\mathcal{E}_{T_1}, S_C]$, over $i \in \{1,\ldots,n\}$ and $t \in \{p,\ldots,T-1\}$. Now,

$$E[K_{b_T}(d_t\,(i)\,,Q)\eta_{i,t+1}\,(s)\,|\mathcal{E}_{T_1}, S_C] = E\left[e^{-D(d_t(i),Q)/b_T}\eta_{i,t+1}\,(s)\,|\mathcal{E}_{T_1}, S_C\right].$$

Writing the expectation in terms of a sum over all possible datacubes, and noting that everything is bounded, gives the following:

$$\begin{aligned}&E\left[e^{-D(d_t(i),Q)/b_T}\eta_{i,t+1}\,(s)\,|\mathcal{E}_{T_1}, S_C\right]\\ &\quad = E[\eta_{i,t+1}\,(s)\,|d_t\,(i) = Q, \mathcal{E}_{T_1}, S_C]P(d_t\,(i) = Q|\mathcal{E}_{T_1}, S_C) + O(e^{-\epsilon/b_T}).\end{aligned}$$

Recalling that $E[\widehat{f}_T(s,Q)|\mathcal{E}_{T_1}, S_C]$ was an average of the above terms, we see that it equals:

$$\frac{1}{n(T-p)}\sum_{t,i}E[\eta_{i,t+1}\,(s)\,|d_t\,(i) = Q, \mathcal{E}_{T_1}, S_C]\cdot P(d_t\,(i) = Q|\mathcal{E}_{T_1}, S_C) + O(e^{-\epsilon/b_T}). \tag{5.1}$$

We will now show that the above average converges to $g(s,Q)R$ for some $R > 0$. The second term in the RHS in eq. (5.1) converges to zero, since $b_T \to 0$ as $T \to \infty$. For the numerator of the first term we have, $E[\eta_{i,t+1}\,(s)\,|d_t\,(i) = Q, \mathcal{E}_{T_1}, S_C]\cdot P(d_t\,(i) = Q|\mathcal{E}_{T_1}, S_C) = \sum_\eta \eta P(\eta_{i,t+1}\,(s) = \eta, d_t\,(i) = Q|\mathcal{E}_{T_1}, S_C)$. Both $d_t\,(i)$ and $\eta_{i,t+1}\,(s)$ are fully determined given the current state S_t of the Markov chain. Using $I_S(X)$ to denote an indicator of X in state S, we have $P(\eta_{i,t+1}\,(s) = \eta, d_t\,(i) = Q|\mathcal{E}_{T_1}, S_C) = \sum_S I_S(\eta_{i,t+1}\,(s) = \eta, d_t\,(i) = Q)P(S_t = S|\mathcal{E}_{T_1}, S_C)$. As a result of this, the first term in the R.H.S of eq (5.1) becomes an average of the form $\frac{1}{T}\sum_t\sum_S \xi(S)P(S_t = S|\mathcal{E}_{T_1}, S_C)$, where $\xi(S) = \frac{1}{n}\sum_{i,\eta}\eta I_S(\eta_{i,t+1}\,(s) = \eta, d_t\,(i) = Q)$. Since we have a finite state-space and $\xi(S)$ is bounded, we can rewrite the above expression as $\sum_S \xi(S)\frac{\sum_t P(S_t = S|\mathcal{E}_{T_1}, S_C)}{T}$.

Now, recall that the query datacube at T is a function of the state S_T, which belongs to a closed irreducible set C with probability 1. Due to stationarity (or cyclic stationarity with a finite cycle length) the average $\sum_t P(S_t = S|\mathcal{E}_{T_1}, S_C)/T$ converges to some constant $R(S)$ (constant because it is a function of the finite state space). For the special case of $S = S_T$, we have the following: (a) $S_T \in C$, so $R(S_T) > 0$, and (b) S_T contains at least one pair of nodes with the feature vector s (since we are attempting link prediction for such a pair), so there exists some $\eta > 0$ for which $I_{S_T}(\eta, Q) = 1$. Together, these imply that $\sum_S \xi(S)\left(\sum_t P(S_t = S|\mathcal{E}_{T_1}, S_C)/T\right)$ converges to some $R_c > 0$, where R_c is a deterministic function of communication class C.

Noting that $E[\widehat{f}_T(s,Q)|S_C] = E[E[\widehat{f}_T(s,Q)|\mathcal{E}_{T_1}, S_C]|S_C]$, and the fact that $\widehat{f}_T$ is bounded we invoke the Dominated Convergence Theorem and see that $E[\widehat{f}_T(s,Q)|S_C] \to R_c$ as well, thus completing the proof of the theorem. □

Lemma 5.4. *Define* $B_T(s,Q,C) = (E[\widehat{h}_T(s,Q)|S_C] - gE[\widehat{f}_T(s,Q)|S_C])/E[\widehat{f}_T(s,Q)|S_C]$. *If Assumption 1 holds, then, we have* $B_T = O(b_T)$. *Since* $b_T \to 0$ *as* $T \to \infty$, *this implies* $B_T = o(1)$.

Proof. For $t \in [p, T-2]; i \in [1,N]; s = \mathrm{s}_T(q,q')$, the numerator of B_T is an average of the terms:

$$A_t := E\left[K_{b_T}(d_t(i), Q)\eta^+_{i,t+1}(s)\,|S_C\right] - E\left[K_{b_T}(d_t(i), Q)\eta_{i,t+1}(s)\,|S_C\right] g(s,Q).$$

Taking expectations w.r.t. $d_t(i)$, and denoting $K_{b_T}(d_t(i), Q)$ by γ, the first term becomes:

$$E\left[\gamma\eta^+_{i,t+1}(s)\,|S_C\right] \quad = \quad E\left[\gamma E\left[\eta^+_{i,t+1}(s)\,|d_t(i), S_C\right]|S_C\right].$$

Now note that $E\left[\eta^+_{i,t+1}(s)\,|d_t(i), S_C\right] = E[E[\eta^+_{i,t+1}(s)\,|d_t(i), \mathcal{E}_{T_1}, S_C]|S_C]$. Conditioning on $\mathcal{E}_{T_1}$ makes $\eta^+_{i,t+1}(s)$ conditionally independent of S_C given $d_t(i)$ if $t > T_1$. Also, for $t \geq T_1$, $E\left[\eta^+_{i,t+1}(s)\,|d_t(i), \mathcal{E}_{T_1}, S_C\right] = \eta_{i,t+1}(s) \cdot g(s, d_t(i))$, as can be seen by summing Eq. 2.3 over all pairs (i,j) in a neighborhood with identical $\mathrm{s}_t(i,j)$, and then taking expectations[2]. This along with the fact that $\gamma\eta^+_{i,t+1}(s)$ is bounded leads to:

$$\begin{aligned} E[\eta^+_{i,t+1}(s)\,|d_t(i), \mathcal{E}_{T_1}, S_C] &\leq \eta_{i,t+1}(s)\,g(s,d_t(i))\mathbf{1}[T_1 \leq t] + c\mathbf{1}[T_1 > t] \\ &\leq \eta_{i,t+1}(s)\,g(s,d_t(i)) + c\mathbf{1}[T_1 > t]. \end{aligned}$$

Thus the numerator of B_T can be upper bounded as:

$$\left|\sum_t A_t/T\right| \leq \sum_t |E[\gamma\eta_{i,t+1}(s)\,(g(s,d_t(i)) - g(s,Q))|S_C]|/T + c'\sum_t P[T_1 > t]/T.$$

The second part is simply $O(E[T_1]/T)$ and $o(1)$. Thus, the numerator of B_T becomes an average of the terms of the following form:

$$E\left[K_{b_T}(d_t(i), Q)\eta_{i,t+1}(s) \cdot (g(s,d_t(i)) - g(s,Q))\,|S_C\right].$$

This expectation is over all possible configurations of the neighborhoods $N_t(i)$ and $N_{t+1}(i)$. Since our neighborhood sizes are bounded (because n is bounded), the expectation is a sum over a finite number of terms.

We now use the smoothness assumption on g. Using $|g(s,d_t(i)) - g(s,Q)| = O(D(d_t(i), Q))$ and that $\eta_{i,t+1}(s)$ is finite for all T and Lemma 5.3, we have:

$$B_T = O\left(E[D(d_t(i), Q)e^{-D(d_t(i),Q)/b_T}|S_C]\right) = O(b_T).$$

The last equation holds since for non-negative x, $xe^{-x/b_T} \leq b_T/e$. □

[2] Note that the conditioning on $\mathcal{E}_{T_1}$ is crucial here.

Lemma 5.5. *Consider an irreducible and aperiodic finite state Markov chain with probability transition matrix P, starting distribution π_0 and stationary distribution π. Let X_t be a deterministic function (with finite support) of the state at time t. The expectation of X under the distribution at time t is denoted by $E[X_t|\pi_0]$. Let μ denote the expectation of X_∞ (i.e. under distribution π). There exists a constant $\lambda \in (0,1)$, and a constant M such that, $\forall t > M$, $\max_{x\in\mathcal{S}} \sum_{y\in\mathcal{S}} |P^t(x,y) - \pi(y)| = O(\lambda^t)$, and $|E[X_t|\pi_0] - \mu| = O(\lambda^t)$.*

Proof. Using the same line of reasoning as [11], we first prove the above for $\max_x \sum_{y\in\mathcal{S}} |P^t(x,y) - \pi(y)|$. Here $|\mathcal{S}|$ denotes the state space and P the $|\mathcal{S}|\times|\mathcal{S}|$ probability transition matrix associated with the Markov chain. Denote by Π the matrix $\mathbf{1}\pi^T$, where $\mathbf{1}$ denotes the column vector of all ones. Note that since $P\Pi = \Pi$ and $\Pi P = \Pi$, we have $P^t - \Pi = (P - \Pi)^t$. For a finite state space irreducible and aperiodic Markov chain, $|P^t(x,y) - \Pi(x,y)| \to 0$, as $t \to \infty$. Hence for some positive $\delta < 1$, we can find an M s.t. $\forall t > M$, $\sum_y |P^t(x,y) - \Pi(x,y)| \le \delta$, $\forall x \in \mathcal{S}$. Since $\max_x \sum_y |P^t(x,y) - \Pi(x,y)| = ||P^t - \Pi||_\infty$, using matrix norm inequalities we have for $t = kM + \ell$, where $\ell < M$ and $t > M$,

$$|P^t - \Pi|_\infty \le ||P^M - \Pi||_\infty^k ||P^\ell - \Pi||_\infty = O(\delta^k),$$

since $\max_{\ell\le M} ||P^\ell - \Pi||_\infty$ is a constant. However, $\delta^k = \delta^{k+1}/\delta = O(\lambda^t)$, where $\lambda = \delta^{1/M} < 1$. Now for $t > M$ and $\lambda < 1$, we have:

$$\max_x |P^t(x,y) - \pi(y)| = O(\lambda^t).$$

First consider π_0 to be an atom at a state $x_0 \in \mathcal{S}$. Since $|E(X_t|X_0) - \mu| \le \sum_{x\in\mathcal{S}} |x||P(x_0,x) - \pi(x)|$, using that X_t is bounded we have the main result. The result can be easily extended to the more general case where π_0 is a convex combination of atoms at $x \in \mathcal{S}$. □

Lemma 5.7. *For any finite integer k, we have*

$$\mathrm{var}\Bigg(\sum_{t\ge T_1+M} q_t|\mathcal{E}_{T_1}, T_1 = k, S_C\Bigg)/T \to \sigma_c \qquad \textit{for some } \sigma_c \ge 0 \tag{5.7}$$

$$\mathrm{var}\Bigg(\sum_t q_t|\mathcal{E}_{T_1}, T_1 = k, S_C\Bigg)/T \to \sigma_c \qquad \textit{for some } \sigma_c \ge 0. \tag{5.8}$$

For a finite state space Markov chain, we also have $E[\mathrm{var}(U_T|\mathcal{E}_{T_1}, S_C)|S_C] \to \sigma_c$ for some $\sigma_c \ge 0$.

Proof. Let C have (finite) period d; the period is finite from the finiteness of the Markov chain, and is typically very small (e.g., $d = 1$ if $0 < g(.) < 1$ everywhere). Let $\mathcal{M}'$ be a Markov chain where each transition corresponds to d transitions of the original chain. Now, $\mathcal{M}'$ is irreducible and aperiodic (since C was irreducible and had period d). Thus, $\exists M, \lambda \in (0,1)$ s.t. $\forall t \ge M$, it is geometrically ergodic with rate λ (Lemma 5.5), which implies in turn that for $t \ge M$, $\mathcal{M}'$ is strongly mixing with exponential drop-off [21] for large k: $\alpha(k) \sim e^{-\beta k}$ for some $\beta > 0$.

Thus, distant states are almost independent, and we use this to bound the covariances of the q_{it}, as follows. Also define $q_t = \sum_i q_{it}/n$.

For the first term, we have:

$$
\begin{aligned}
&(1/T)\mathrm{var}\left[\sum_{t=1}^{T} q_t | \mathcal{E}_{T_1}, S_C\right] \\
&= \underbrace{(1/T)\sum_{t<T_1,t'<T_1} \mathrm{cov}(q_t, q_{t'}|\mathcal{E}_{T_1}, S_C)}_{(P_0)} + \underbrace{(1/T)\sum_{t\geq T_1} \mathrm{var}(q_t|\mathcal{E}_{T_1}, S_C)}_{(P_1)} \\
&+ \underbrace{(2/T)\sum_{t<T_1,t'\geq T_1} \mathrm{cov}(q_t, q_{t'}|\mathcal{E}_{T_1}, S_C)}_{(P_2)}.
\end{aligned}
$$

First, note that $P_0 = O(T_1^2/T)$. We now focus on P_1. Let $U := \sum_{T_1 \leq t < T_1+M} q_t$, and $V := \sum_{t\geq T_1+M} q_t$. Thus,

$$
\mathrm{var}\left(\sum_{t\geq T_1} q_t|\mathcal{E}_{T_1}, S_C\right) = \mathrm{var}(U|\mathcal{E}_{T_1}, S_C) + \mathrm{var}(V|\mathcal{E}_{T_1}, S_C) + \mathrm{cov}(U, V|\mathcal{E}_{T_1}, S_C).
$$

$\mathrm{var}(U|\mathcal{E}_{T_1}, S_C) = O(M^2)$, as for $\mathrm{var}(V|\mathcal{E}_{T_1}, S_C)$, we have:

$$
\mathrm{var}(V|\mathcal{E}_{T_1}, S_C) = (2/T)\sum_{t\geq T_1+M} \underbrace{\sum_{t'\geq t} \mathrm{cov}(q_t, q_{t'}|\mathcal{E}_{T_1}, S_C)}_{A_t}.
$$

Recall that for our Markov chain, S_t involves $p+1$ graphs $(G_{t-p+1}, \ldots, G_{t+1})$. Since q_t is a function of S_t, it also depends on $p+1$ graphs. Hence, the distance $\mathrm{dist}(t, t')$ between two sigma-algebras $\mathcal{F}_{\leq t}$ and $\mathcal{F}_{>t'}$ is defined as $\max(\lceil (t' - t - (p+1))/d \rceil, 0)$. Thus, the total number of states at distance k is $O(1)$. Let $R_t = \lfloor (T-t)/d \rfloor$. Rather importantly, note that we will use basic conditional independence results from Markov chains. For example $E[X_t X_{t+2d}|X_{t+d}] = E[X_t|X_{t+d}]E[X_{t+2d}|X_{t+d}]$. Unfortunately, conditioned on $\mathcal{E}_{T_1} \cap S_C$ this may not be true. However, if $t \geq T_1$, we can safely use the conditional independence, which is definitely true for A_t.

For notational convenience we will denote by cov_c and E_c covariance and expectation conditioned on $\mathcal{E}_{T_1} \cap S_C$. Then,

$$
\begin{aligned}
A_t &= \sum_{t\leq t'<t+(R_t-1)d} \mathrm{cov}_c(q_t, q_{t'}) + \sum_{t+R_t d\leq t'\leq T} \mathrm{cov}_c(q_t, q_{t'}) \\
&= \sum_{r=0}^{R_t-1}\sum_{\ell=0}^{d-1} \mathrm{cov}_c(q_t, q_{t+rd+\ell}) + \sum_{t+R_t d\leq t'\leq T} \mathrm{cov}_c(q_t, q_{t'}) \\
&= \sum_r (E_c[q_t u_{tr}] - E_c[q_t]E_c[u_{tr}]) + (E_c[q_t u_{tR_t}] - E_c[q_t]E_c[u_{tR_t}])
\end{aligned}
$$

$$\begin{aligned}
&\left(\text{letting } u_{tr} = \sum_{\ell=0}^{d-1} q_{t+rd+\ell} \text{ and } u_{tR_t} = \sum_{t'\geq t+R_t d}^{T} q_{t+rd+\ell}\right)\\
&= \sum_r \left(E_c[E_c[q_t u_{tr} \mid S'_{t+rd}]] - E_c[q_t]E_c[u_{tr}]\right)\\
&\quad + \left(E_c[E_c[q_t u_{tR_t} \mid S'_{t+R_td}]] - E_c[q_t]E_c[u_{tR_t}]\right)\\
&= \sum_r \left(E_c[E_c[q_t \mid S'_{t+rd}]E_c[u_{tr} \mid S'_{t+rd}]] - E_c[q_t]E_c[E_c[u_{tr} \mid S'_{t+rd}]]\right)\\
&\quad + \left(E_c[E_c[q_t u_{tR_t} \mid S'_{t+R_td}]] - E_c[q_t]E_c[u_{tR_t}]\right) \quad \text{By Markov property}\\
&= \sum_r \left(E_c[E_c[q_t \mid S'_{t+rd}]p(S'_{t+rd})] - E_c[q_t]E_c[p(S'_{t+rd})]\right)\\
&\quad + \left(E_c[E_c[q_t \mid S'_{t+R_td}]p(S'_{t+R_td})] - E_c[q_t]E_c[p(S'_{t+R_td})]\right)\\
&\quad\ E_c[u_{tr} \mid S'_{t+rd}] \text{ is denoted as a function } p(.)\\
&= \sum_r \left(E_c[E_c[q_t p(S'_{t+rd}) \mid S'_{t+rd}]] - E_c[q_t]E_c[p(S'_{t+rd})]\right)\\
&\quad + \left(E_c[E_c[q_t p(S'_{t+R_td}) \mid S'_{t+R_td}]] - E_c[q_t]E_c[p(S'_{t+R_td})]\right)\\
&= \sum_r \left(E_c[q_t p(S'_{t+rd})] - E_c[q_t]E_c[p(S'_{t+rd})]\right)\\
&\quad + \left(E_c[q_t p(S'_{t+R_td})] - E_c[q_t]E_c[p(S'_{t+R_td})]\right)\\
&= B_t + \operatorname{cov}_c(q_t, p(S'_{t+R_td})) \quad \text{where } B_t = \sum_r \operatorname{cov}_c(q_t, p(S'_{t+rd})).
\end{aligned}$$

Recall that we were originally interested in $\sum_{t>T_1} A_t/T$. Let us first consider $1/T \sum_t \operatorname{cov}(q_t, p(S'_{t+R_td})|\mathcal{E}_{T_1}, S_C)$. By virtue of geometric ergodicity $\operatorname{cov}(q_t, p(S'_{t+R_td})|\mathcal{E}_{T_1}, S_C) = O\left(e^{-\beta R_t}\right)$, where $R_t = \lfloor (T-t)/d \rfloor$. Thus we have:

$$\sum_t |\operatorname{cov}(q_t, p(S'_{t+R_td})|\mathcal{E}_{T_1}, S_C)| = O\left(\sum_t e^{-\beta\lfloor (T-t)/d\rfloor}\right) = O\left(\frac{e^{\beta}}{1-e^{-\beta/d}}\right).$$

Using elementary arguments from real analysis we see that $\sum_t \operatorname{cov}(q_t, p(S'_{t+R_td})|\mathcal{E}_{T_1}, S_C)$ converges to some finite number. Hence after dividing by T it contributes a $o(1)$ term to the expression $\sum_t A_t/T$. For this reason we will now concentrate on $\sum_{t>T_1} B_t/T$ term. First note that the sequence B_t is upper bounded by the following,

$$\begin{aligned}
B_t &\leq \sum_r |\operatorname{cov}(q_t, p(S'_{t+rd})|\mathcal{E}_{T_1}, S_C)| && \text{Also } t > T_1\text{, and we have conditioned on } \mathcal{E}_{T_1}, S_C\\
&\leq O\left(\sum_r e^{-\beta r}\right) = O(1) && \text{Since all } q_t \text{ are bounded.}
\end{aligned}$$

We again see that B_t also converges to some constant c_t, thus making P_2 asymptotically equivalent to: $\frac{1}{T}\sum_{\ell=0}^{d-1}\sum_{r=0}^{R_t-1} c_{T_1+rd+\ell}$. However, for all $T_1 \leq T$, if the chain is cyclo-stationary, then after a finite time, for any $\ell \in \{0,\ldots,d-1\}$,

$c_{T_1+rd+\ell}$ approaches the same constant c_ℓ, $\forall r$. Therefore, for all $T_1 \leq T$ we have $\lim_{R\to\infty} \sum_{r=0}^{R-1} c_{T_1+rd+\ell}/R = c_\ell$, where c_ℓ is a constant w.r.t T. This leads to:

$$\mathrm{var}(V|\mathcal{E}_{T_1}, S_C) \to 1/d \sum_{\ell=0}^{d-1} c_\ell \qquad \text{as } T \to \infty.$$

Since the P_1 is a variance term, it is non-negative for all T, and hence $\sigma_c = 1/d \sum_{\ell=0}^{d-1} c_\ell$ must be non-negative as well, thus proving Equation 5.7. Using the Cauchy Schwartz inequality,

$$\mathrm{cov}(U, V|\mathcal{E}_{T_1}, S_C)/T = O(\sqrt{(\mathrm{var}(U|\mathcal{E}_{T_1}, S_C)/T)(\mathrm{var}(V|\mathcal{E}_{T_1}, S_C)/T)}) = o(1).$$

Thus $P_1 \to \sigma_c$ as $T \to \infty$ for some non-negative constant σ_c.

Another use of the Cauchy Schwartz argument from before, along with the convergence result on P_1 lets us upper bound P_2 by $O(T_1/\sqrt{T})$.

Thus, for finite k, putting all the bounds (i.e. on P_0, P_1, and P_2) together, we have $\mathrm{var}(\sum_t q_t|\mathcal{E}_{T_1}, S_C, T_1 = k)/T \to \sigma_c$, for some $\sigma_c \geq 0$, proving Equation 5.8. Also, since T_1 has finite first and second moments for a finite space Markov chain, we have $E[\mathrm{var}(\sum_t q_t|\mathcal{E}_{T_1}, S_C)|S_C]/T \to \sigma_c$.

We remind the reader that using simple arguments for finite state space Markov chains, it can be shown that T_1's tail probability is geometrically decaying, leading to the finiteness of the first and second moments. □

Lemma 5.8. $\mathrm{var}(E[U_T|\mathcal{E}_{T_1}, S_C]|S_C) = o(1)$.

Proof. Recall that $U_T := \sum_t q_t/\sqrt{T}$. Let μ_c denotes the expectation of q_t under the stationary distribution in communication class C (it is a deterministic function of class C). Since $\mathrm{var}(E[q_t|\mathcal{E}_{T_1}, S_C]|S_C) = \mathrm{var}(E[q_t|\mathcal{E}_{T_1}, S_C] - \mu_c|S_C)$, we will simply upper bound $E[U_T - \mu_c|\mathcal{E}_{T_1}, S_C]$. Lemma 5.5 shows that: $\exists M$, and $\lambda \in (0, 1)$ such that, $\forall t > T_1 + M$, $|E[q_t|\mathcal{E}_{T_1}, S_C] - \mu_c| = O(\lambda^{t-T_1})$. Thus,

$$|E[U_T - \mu_c|S_C, \mathcal{E}_{T_1}]| \leq \frac{c(T_1 + M)}{\sqrt{T}} + \frac{\sum_{t>T_1+M} \lambda^{t-T_1}}{\sqrt{T}} = O\left(\frac{T_1 + M}{\sqrt{T}}\right) \qquad (5.9)$$

Thus, $\mathrm{var}(E[U_T|\mathcal{E}_{T_1}, S_C]) = O\left(E[(T_1 + M)^2]/T\right) = o(1)$, since T_1 has finite second moment. □

A.2. Statement and proofs of results from Section 6

Lemma 6.2. *The Wasserstein distance $d_W(W, Z)$ between W and the standard normal random variable Z is upper bounded as follows:*

$$d_W(W, Z) \leq \min_{k\leq T} \left(c_1 B^3 \gamma_T \left(\frac{|N_{\leq k}|}{\sigma_T}\right)^2 + c_2 B \gamma_T \alpha(k) \right.$$
$$\left. + B^2 \sqrt{c_3 \left(\frac{\gamma_T \tau_k}{\sigma_T}\right)^2 + c_4 \gamma_T \left(\frac{|N_{\leq k}|}{\sigma_T}\right)^3 + c_5 \gamma_T \frac{\tau_k}{\sigma_T} \left(\frac{|N_{\leq k}|}{\gamma_T}\right)^2} \right),$$

where c_1, c_2, c_3, c_4, c_5 are constants.

Proof. We define the following sets:

$$N_m(i) := \{j : \mathrm{dist}(i,j) = m\}, \quad N_{\le k}(i) := \bigcup_{m \le k} N_m(i), \quad N_{>k}(i) := \bigcup_{m>k} N_m(i).$$

We also define the following upper bounds on the sizes of these sets:

$$|N_m| := \max_i |N_m(i)|, \quad |N_{\le k}| := \max_i |N_{\le k}(i)|, \quad |N_{>k}| := \max_i |N_{>k}(i)|.$$

Before beginning, we recall two facts.

(1) *Bounded covariance via strong mixing:* For two random variables X and Y that are more than distance k away, we have

$$|E[XY] - E[X]E[Y]| \le 4\|X\|_\infty \|Y\|_\infty \alpha(k).$$

(2) *Bounds on Wasserstein distance:* For the set of functions $\mathcal{F} = \{f \mid \|f\|, \|f''\| \le 2, \|f'\| \le \sqrt{2/\pi}\}$,

$$d_W(W, Z) \le \sup_{f \in \mathcal{F}} |E[f'(W) - Wf(W)]|,$$

where $d_W(.)$ is the Wasserstein distance and Z has the standard normal distribution.

In the following, we shall bound $|E[f'(W) - Wf(W)]|$. We shall repeatedly break up W into two parts: $W_i = \sum_{j \in N_{>k}(i)} X_j$ being the contribution from all nodes within a distance k of some node i, and the remainder from nodes "far away" from i. Here, k is a parameter that shall be picked later. We can bound $|E[f'(W) - Wf(W)]|$ as follows:

$$\begin{aligned} |E[f'(W) - Wf(W)]| &= \left| E\left[f'(W) - \sum_i X_i f(W) \right] \right| \\ &\le \left| E\left[f'(W) \left(1 + \sum_i X_i (W_i - W) \right) \right] \right| \\ &\quad + \left| E\left[\sum_i X_i (W_i - W) f'(W) + \sum_i X_i f(W) \right] \right|. \end{aligned} \tag{6.10}$$

The second part in eq. 6.10 can be further bounded above as follows,

$$\begin{aligned} &\left| E\left[\sum_i X_i (W_i - W) f'(W) + \sum_i X_i f(W) \right] \right| \\ &\le E\left| \sum_i X_i (W_i - W) f'(W) - \sum_i X_i (f(W_i) - f(W)) \right| + \left| E\left[\sum_i X_i f(W_i) \right] \right| \\ &\le \frac{1}{2} E\left| \sum_i X_i (W - W_i)^2 f''(W_i^*) \right| + \left| E\left[\sum_i X_i f(W_i) \right] \right| \end{aligned} \tag{6.11}$$

$$\leq \frac{\|f''\|}{2} E \left| \sum_i X_i (W_i - W)^2 \right| + \left| E\left[\sum_i X_i f(W_i) \right] \right| ,$$

where the second inequality follows from Taylor expansion with W_i^* being some value between W and W_i.

First, note that:

$$\begin{aligned}
\|f''\| E \left| \sum_i X_i (W_i - W)^2 \right| &= \|f''\| E \left| \sum_i \sum_{j1, j2 \in N_{\leq k}(i)} X_i X_{j1} X_{j2} \right| \\
&\leq \|f''\| \sum_i \sum_{j1, j2 \in N_{\leq k}(i)} E|X_i X_{j1} X_{j2}| \\
&\leq \|f''\| \sum_i \sum_{j1, j2 \in N_{\leq k}(i)} \frac{E|X_i^3| + E|X_{j1}^3| + E|X_{j2}^3|}{3} \\
&\leq 2 c_1 B^3 \frac{T |N_{\leq k}|^2}{{\sigma_T}^3}
\end{aligned}$$

(The factor 2 is added for later ease of notation). As for the second term in eq. 6.11 we have:

$$\begin{aligned}
\left| E\left[\sum_i X_i f(W_i) \right] \right| &\leq \sum_i |E[X_i f(W_i) - E[X_i] E[f(W_i)]]| \quad \text{(because } E[X_i] = 0\text{)} \\
&= \sum_i |\text{cov}(X_i, f(W_i))| \leq \frac{4\|f\| B T \alpha(k)}{\sigma_T} = c_2 B \frac{T \alpha(k)}{\sigma_T}.
\end{aligned}$$

Thus, we obtain a bound for both terms in eq. 6.11, and hence a bound for the second term of eq. 6.10. We will now bound the first term in eq. 6.10. Let $P_T = \sum_i X_i (W_i - W)$. Denote by τ_k the tail sum $\sum_{m>k} |N_m| \alpha(m)$. Recall that $E[X_i] = 0$ and $E[W^2] = 1$. Thus,

$$\begin{aligned}
\left| E\left[f'(W) \left(1 + \sum_i X_i (W_i - W) \right) \right] \right| &\leq E\left| f'(W) \left(1 + P_T \right) \right| \leq \|f'\| \sqrt{E\left[1 + P_T\right]^2} \\
&\leq \|f'\| \sqrt{E[(1 + E[P_T]) + (P_T - E[P_T])]^2} \\
&\leq \sqrt{2/\pi} \sqrt{(1 + E[P_T])^2 + \text{var}(P_T)}.
\end{aligned}$$

Now,

$$\begin{aligned}
|E[P_T] + 1| = \left| E\left[\sum_i X_i W_i \right] \right| &= \left| \sum_i E\left[X_i \sum_{j \in N_{>k}(i)} X_j \right] \right| \\
&= \left| \sum_i \sum_{m>k} \sum_{j \in N_m(i)} E[X_i X_j] \right|
\end{aligned}$$

$$= \left|\sum_i \sum_{m>k} \sum_{j\in N_m(i)} (E[X_iX_j] - E[X_i]E[X_j])\right|$$
$$\leq \sum_i \sum_{m>k} \frac{c''B^2}{{\sigma_T}^2}\alpha(m)|N_m| \leq c''B^2\frac{T\tau_k}{{\sigma_T}^2}.$$

Next, we look at the var(P_T) term:

$$\begin{aligned}\mathrm{var}(P_T) &= (E[P_T^2] - E[P_T]^2) \qquad (6.12)\\ &= E\left[\left(\sum_{\substack{i\\ j\in N_{\leq k}(i)}} X_iX_j\right)^2\right] - E[P_T]^2\\ &= \underbrace{E\left[\sum_{\substack{i,j\\ s\in N_{\leq k}(i)\\ t\in N_{\leq k}(j)}} X_iX_jX_sX_t\right]}_{(A)} - E[P_T]^2.\end{aligned}$$

The first term (i.e., term (A)) in eq. 6.12 can be broken into two parts, one such that the minimum distance between any node in $\{i, s\}$ and any node in pair $\{j, t\}$ is $\leq k$ (denote this by set $F_{\leq k}$), and one where its greater than k (denote this by set $F_{>k}$). Formally, we define the following terms:

$$F_m = \{(i, j, s, t) : s \in N_{\leq k}(i), t \in N_{\leq k}(j), \min_{a,b\in\{i,j,s,t\}} \mathrm{dist}(a, b) = m\}$$
$$F_{\leq k} = \bigcup_{m\leq k} F_m, \quad F_{>k} = \bigcup_{m>k} F_m, \quad |F_m| = \max_i |F_m(i)|, \quad |F_{\leq k}| = \max_i |F_{\leq k}(i)|.$$

Consider the term $|F_{\leq k}|$. Given i, s can be picked in at most $|N_{\leq k}|$ ways. Now, either j or t or both must be within distance k of i or s. Thus, given i and s, j (or t) can be picked in at most $2|N_{\leq k}|$ ways, and then t (or j) can be picked in another $|N_{\leq k}|$ ways. Hence, $|F_{\leq k}| \leq 4T|N_{\leq k}|^3$. By a similar argument, $|F_m| \leq 4T|N_{\leq k}|^2|N_m|$.

Now, we have:

$$\begin{aligned}(A) &= \sum_{F_{\leq k}} E[X_iX_jX_sX_t] + \sum_{F_{>k}} E[X_iX_jX_sX_t]\\ &= \sum_{F_{\leq k}} E[X_iX_jX_sX_t] + \sum_{F_{>k}} E[X_iX_s]E[X_jX_t]\\ &\quad + \sum_{F_{>k}} (E[X_iX_jX_sX_t] - E[X_iX_s][X_jX_t])\end{aligned}$$

$$\leq \underbrace{\sum_{F_{\leq k}} E[X_i X_j X_s X_t]}_{(B0)} + \underbrace{\sum_{F_{>k}} E[X_i X_s] E[X_j X_t]}_{(B1)} + \underbrace{4 \sum_{m>k} \sum_{F_m} \frac{B^4}{{\sigma_T}^4} \alpha(m)}_{(B2)} .$$

$$\begin{aligned}
(B0) &= \sum_{F_{\leq k}} E[X_i X_j X_s X_t] \leq \sum_{F_{\leq k}} \frac{E[X_i^4] + E[X_j^4] + E[X_s^4] + E[X_t^4]}{4} \\
&\leq \frac{B^4}{{\sigma_T}^4} \sum_{F_{\leq k}} 1 \leq 4B^4 \frac{T|N_{\leq k}|^3}{{\sigma_T}^4}. \\
(B1) &= \sum_{F_{>k}} E[X_i X_s] E[X_j X_t] \\
&= \sum_{F_{>k} \bigcup F_{\leq k}} E[X_i X_s] E[X_j X_t] - \sum_{F_{\leq k}} E[X_i X_s] E[X_j X_t] \\
&\leq \left(\sum_i E[X_i (W - W_i)] \right)^2 + \sum_{F_{\leq k}} \frac{E[X_i^4] + E[X_s^4] + E[X_j^4] + E[X_t^4]}{4} \\
&\leq (E[P_T])^2 + 4B^4 \frac{T|N_{\leq k}|^3}{{\sigma_T}^4}. \\
(B2) &\leq \frac{4B^4}{{\sigma_T}^4} \sum_{m>k} |F_m| \alpha(m) \leq 16B^4 \frac{T|N_{\leq k}|^2}{{\sigma_T}^4} \sum_{m>k} |N_m| \alpha(m) \leq 16B^4 \frac{T|N_{\leq k}|^2}{{\sigma_T}^4} \tau_k.
\end{aligned}$$

The last equation simply uses a number of applications of the fact that the geometric mean is less than the arithmetic mean, and Jensen's inequality. Plugging these into Equation 6.12, we have:

$$\begin{aligned}
\operatorname{var}(P_T) &= (B0) + (B1) + (B2) - E[P_T]^2 \\
&\leq 4B^4 \frac{T|N_{\leq k}|^2}{{\sigma_T}^4} + 4B^4 \frac{T|N_{\leq k}|^3}{{\sigma_T}^4} + 16B^4 \frac{T|N_{\leq k}|^2}{{\sigma_T}^4} \tau_k \\
&\leq 8B^4 \frac{T|N_{\leq k}|^3}{{\sigma_T}^4} + 16B^4 \frac{T|N_{\leq k}|^2}{{\sigma_T}^4} \tau_k.
\end{aligned}$$

Combining these steps, and recalling that $\gamma_T = T/\sigma_T$, we finally obtain the following form for Eq. 6.10:

$$\begin{aligned}
d_W(W, Z) \leq c_1 B^3 \frac{T|N_{\leq k}|^2}{{\sigma_T}^3} + c_2 B \frac{T\alpha(k)}{\sigma_T} \\
+ \|f'\| \sqrt{c''^2 B^4 \frac{T^2 \tau_k^2}{{\sigma_T}^4} + 8B^4 \frac{T|N_{\leq k}|^3}{{\sigma_T}^4} + 16B^4 \frac{T|N_{\leq k}|^2}{{\sigma_T}^4} \tau_k} \\
\leq c_1 B^3 \gamma_T \left(\frac{|N_{\leq k}|}{\sigma_T} \right)^2 + c_2 B \gamma_T \alpha(k)
\end{aligned}$$

$$+ B^2 \sqrt{c_3 \left(\frac{\gamma_T \tau_k}{\sigma_T}\right)^2 + c_4 \gamma_T \left(\frac{|N_{\leq k}|}{\sigma_T}\right)^3 + c_5 \gamma_T \left(\frac{|N_{\leq k}|}{\sigma_T}\right)^2 \frac{\tau_k}{\sigma_T}}.$$

□

A.3. Statement and proofs of results from Section 7

We will start by reminding the reader some of the definitions. Define the following:

$$\begin{aligned}
\widehat{h}_T(t) &:= \frac{1}{n}\sum_{i=1}^{n} K_{b_T}(d_t(i), Q)\eta^+_{i,t+1}(s) \\
\widehat{f}_T(t) &:= \frac{1}{n}\sum_{i=1}^{n} K_{b_T}(d_t(i), Q)\eta_{i,t+1}(s) \\
q_t &:= \widehat{h}_T(t) - E[\widehat{h}_T(t)|S_C] - g(\widehat{f}_T(t) - E[\widehat{f}_T(t)|S_C]) \\
p_t &:= [\widehat{h}_T(t) - g\widehat{f}_T(t)] - E[\widehat{h}_T(t) - g\widehat{f}_T(t)|\mathcal{E}_{T_1}, S_C].
\end{aligned}$$

We define: ${\sigma^2}_T(T_1, C) := \text{var}(\sum_t q_t|\mathcal{E}_{T_1}, S_C)$, and ${\sigma^2}_T(C) := \text{var}(\sum_t q_t|S_C)$. Also, ${\sigma^2}_T(T_1, C) := \text{var}(\sum_t q_t|\mathcal{E}_{T_1}, S_C)$.

Lemma 7.2. *Under Assumption 1 and assuming $\sigma_c > 0$,*

$$\textit{Conditioned on } S_C, \qquad \sum_t q_t/\sqrt{T} \xrightarrow{d} \mathcal{N}(0, \sigma_c^2) \qquad \textit{As } T \to \infty.$$

Proof. Using our distributional convergence results conditioned on $\mathcal{E}_{T_1} \cap S_C$, we have shown that

$$\sum_{t \geq T_1+M} p_t/\sqrt{T} \xrightarrow{d} \mathcal{N}(0, \sigma_c^2) \quad \text{Conditioned on } \mathcal{E}_{T_1} \cap S_C\text{, when } T_1 \text{ has a finite value.}$$

Denote by $V_t := \widehat{h}_T(t) - g\widehat{f}_T(t)$. We have,

$$\begin{aligned}
&\left|\sum_t q_t/\sqrt{T} - \sum_{t \geq T_1+M} p_t/\sqrt{T}\right| \qquad (7.13)\\
&\leq \left|\sum_{t < T_1+M} q_t/\sqrt{T}\right| + \sum_{t \geq T_1+M} |E[V_t|\mathcal{E}_{T_1}, S_C] - E[V_t|S_C]|/\sqrt{T} \\
&\leq c(T_1 + M)/\sqrt{T} + c' \sum_{t \geq T_1+M} \lambda^{t-T_1}/\sqrt{T} \\
&= c''(T_1 + M)/\sqrt{T} \quad \text{Using Lemma 5.5.}
\end{aligned}$$

where c, c' and c'' are positive constants. Let $F_k(x)$ denote the c.d.f of $\sum_{t \geq T_1+M} p_t/\sqrt{T}$, i.e. $F_k(x) = P(\sum_{t \geq k+M} p_t/\sqrt{T} \leq x|\mathcal{E}_{T_1}, S_C, T_1 = k)$. Lemma 7.3 tells us that, for finite k and $\forall x \in \mathcal{R}$, $F_k(x) \to \Phi_{0,\sigma_c^2}(x)$; $\Phi_{0,\sigma_c^2}(x)$ being the c.d.f of a normal distribution with mean zero, and standard deviation σ_c. Now, using Equation 7.13 we have the following simple argument:

$$
\begin{aligned}
&P\left(\sum_t q_t/\sqrt{T} \leq x|S_C\right)\\
&\leq \sum_k P\left(\sum_{t\geq k+M} p_t/\sqrt{T} \leq x + c''(k+M)/\sqrt{T}|\mathcal{E}_{T_1}, S_C, T_1 = k\right) P(T_1 = k|S_C)\\
&\leq \sum_{k\leq K} F_k(x + c''(k+M)/\sqrt{T}) P(T_1 = k|S_C) + P(T_1 > K) \qquad \text{For any finite } K\\
&\rightarrow \limsup_{T\rightarrow\infty} P\left(\sum_t q_t/\sqrt{T} \leq x|S_C\right) \leq \Phi_{0,\sigma_c^2}(x) P(T_1 \leq K) + P(T_1 > K).
\end{aligned}
$$

In the last step, the exchange of limit and expectation is valid by virtue of the Dominated Convergence Theorem. Now taking $K \rightarrow \infty$ (which minimizes the upper bound on the lim sup) and using the geometric bound on tail probability of T_1 in finite state space Markov chains, we have:

$$\limsup_{T\rightarrow\infty} P\left(\sum_t q_t/\sqrt{T} \leq x|S_C\right) \leq \Phi_{0,\sigma_c^2}(x).$$

An identical argument on $P(\sum_t q_t/\sqrt{T} > x|S_C)$ gives the following equation.

$$\liminf_{T\rightarrow\infty} P\left(\sum_t q_t/\sqrt{T} \leq x|S_C\right) \geq \Phi_{0,\sigma_c^2}(x).$$

Thus we show that $\forall x \in \mathcal{R}$, as

$$T \rightarrow \infty \qquad P\left(\sum_t q_t/\sqrt{T} \leq x|S_C\right) \rightarrow \Phi_{0,\sigma_c^2}(x),$$

which in turn proves our result. □

Lemma 7.3. *Define* $p_t := [\widehat{h}_T(t) - g\widehat{f}_T(t)] - E[\widehat{h}_T(t) - g\widehat{f}_T(t)|\mathcal{E}_{T_1}, S_C]$. *Under Assumption 1 and assuming* $\sigma_c > 0$, *for any finite* T_1, *we have:*

$$\sum_{t\geq T_1+M} p_t/\sqrt{T} \xrightarrow{d} \mathcal{N}(0, \sigma_c^2) \qquad \textit{conditioned on } \mathcal{E}_{T_1} \cap S_C \textit{ as } T \rightarrow \infty.$$

Proof. If we can show that the conditions in Lemma 6.3 are satisfied for

$$W_T := \left(\sum_{t\geq T_1+M} p_t\right) \Bigg/ \sqrt{\operatorname{var}\left(\sum_{t\geq T_1+M} p_t|\mathcal{E}_{T_1}, S_C\right)},$$

then using Lemma 6.1 we will have:

$$W_T \xrightarrow{d} \mathcal{N}(0, 1) \qquad \text{conditioned on } \mathcal{E}_{T_1} \cap S_C.$$

However, for any finite value of T_1, $\mathrm{var}(\sum_{t\geq T_1+M} p_t|\mathcal{E}_{T_1}, S_C)/T \rightarrow \sigma_c^2$ (see Lemma 5.7, eq. 5.7). Thus, with the additional assumption of $\sigma_c > 0$, the result is proved.

Now we will show that, conditioned on $\mathcal{E}_{T_1} \cap S_C$, the conditions in Lemma 6.3 are satisfied for W_T, and thus the Wasserstein distance in Lemma 6.2 can be upper bounded by $O(T^{-1/2}\log^2(T))$.

First note that p_t is bounded and $E[p_t|\mathcal{E}_{T_1}, S_C] = 0$. Thus p_t corresponds to Y_t in Lemma 6.2. Since p_t is a function of S_t, it involves $p+1$ graphs $(G_{t-p+1},\dots,G_{t+1})$. The distance $\mathrm{dist}(i,j)$ is defined as $\max(|i-j|-(p+1),0)$. Thus $|N_m|$ equals 2 for $m > 0$, and $2(p+1)$ otherwise; hence $|N_m| = O(1)$. Also, $|N_{\leq k}| = O(k)$. Denote by $\sigma_T(T_1, C)$ the standard deviation of $\sum_{t\geq T_1} p_t$ conditioned on $S_C \cap \mathcal{E}_{T_1}$. Let us now examine the conditions in Lemma 6.3.

Condition 1 $\boxed{\gamma_T \rightarrow \infty}$ We have $\gamma_T = T/\sigma_T(T_1, C) \rightarrow \sqrt{T}/\sigma_c \rightarrow \infty$, where the limits follow from Lemma 5.7 and the $\sigma_c > 0$ assumption.

Condition 2 Let $k(T) = \log T/\beta$. We will show that this satisfies conditions 2a, 2b, and 2c.

2a: $\boxed{\gamma_T\alpha(k(T)) \rightarrow 0}$

Plugging in the value of $k(T)$, and using Lemma 5.7 we see that:

$$\gamma_T\alpha(k(T)) = O\left(Te^{-\beta k(T)}\Big/\sigma_T(T_1, C)\right) = O(T^{-1/2}).$$

2b: $\boxed{\gamma_T\tau_{k(T)}\big/\sigma_T(T_1, C) \rightarrow 0}$

Using Lemma 5.7 we see that:

$$\begin{aligned}
&\gamma_T\tau_{k(T)}\big/\sigma_T(T_1, C) = (T/\sigma_T(T_1, C)^2)\tau_{k(T)} = O(\tau_{k(T)})\\
&= O\Bigg(\sum_{m>k(T)} |N_m|\alpha(m)\Bigg)\\
&= O\Bigg(e^{-\beta k(T)}\sum_{t>0} e^{-\beta t}\Bigg) \quad \text{Using } |N_m| = O(1) \text{ and } \alpha(k) = O(e^{-\beta k}).\\
&= O(e^{-\beta k(T)}) = O(T^{-1}) \quad \text{Using } k(T) = \log T/\beta.
\end{aligned}$$

2c: $\boxed{\gamma_T\left(\frac{|N_{\leq k(T)}|}{\sigma_T(T_1,C)}\right)^2 \rightarrow 0}$

Again, using Lemma 5.7 gives us:

$$\begin{aligned}
&\gamma_T\left(\frac{|N_{\leq k(T)}|}{\sigma_T(T_1, C)}\right)^2 = T/\sigma_T(T_1, C)^2\ |N_{\leq k(T)}|^2\big/\sqrt{T}\\
&= O(|N_{\leq k(T)}|^2\big/\sqrt{T})\\
&= O(k(T)^2\big/\sqrt{T}) = O((\log T)^2/T^{1/2}) \quad \text{Using } k(T) = \log T/\beta.
\end{aligned}$$

Now the upper bound on Wasserstein distance (Lemma 6.2) becomes $O(\log(T)^2/T)$ by using $k = \log(T)/T$ and the expressions derived before as part of the second condition. □

Acknowledgments

We are grateful to Peter J. Bickel for helpful discussions on this topic.

References

[1] Adamic, L. and Adar, E., Friends and neighbors on the web. *Social Networks*, 25:211–230, 2003.

[2] Aitchison, J. and Aitken, C. G. G., Multivariate binary discrimination by the kernel method. *Biometrika*, 63:413–420, 1976. MR0443222

[3] Bradley, R. C., Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107–144, 2005. MR2178042

[4] Chakrabarti, D., Faloutsos, C., and Zhan, Y., Visualization of large networks with min-cut plots, a-plots and r-mat. *Int. J. Hum.-Comput. Stud.*, 65(5):434–445, 2007.

[5] Chen, L. H. Y., Goldstein, L., and Shao, Q. M., *Normal Approximation by Stein's Method*. Springer Verlag, 2010.

[6] Durrett, R., *Probability: Theory and Examples*. Duxbury Press, 1995. MR2722836

[7] Levina, E. and Bickel, P. J., Thexture synthesis and nonparametric resampling of random fields. *Annals of Statistics*, 34(4):1751–1773, 2006. MR2283716

[8] Fu, W., Xing, E. P., and Song, L., A state-space mixed membership blockmodel for dynamic network tomography. *Annals of Applied Statistics*, 4:535–566, 2010. MR2758639

[9] Grimmett, G. and Stirzaker, D., *Probability and Random Processes*. Oxford University Press, 2001. MR2059709

[10] Hanneke, S. and Xing, E. P., Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4:585–605, 2006. MR2660534

[11] Heidergott, B., Hordijk, A., and van Uitert, M., Series expansions for finite-state Markov chains. Tinbergen Institute Discussion Papers 05-086/4, 2005.

[12] Hoff, P. D., Latent factor models for relational data. URL `http://www.stat.washington.edu/hoff/public/acms.pdf`.

[13] Holland, P. W. and Leinhardt, S., A dynamic model for social networks. *Journal of Mathematical Sociology*, 5:5–20, 1977. MR0446596

[14] Huang, Z. and Lin, D. K. J., The time-series link prediction problem with applications in communication surveillance. *INFORMS Journal on Computing*, 2009.

[15] Indyk, P. and Motwani, R., Approximate nearest neighbors: Towards removing the curse of dimensionality. In *ACM Symposium on Theory of Computing*. MIT Press, 1998. MR1715608

[16] Katz, L., A new status index derived from sociometric analysis. In *Psychometrika*, volume 18, pages 39–43, 1953.

[17] Kolar, M., Song, L., Ahmed, A., and Xing, E., Estimating time-varying networks. *Annals of Applied Statistics*, 2010. MR2758086

[18] Liben-Nowell, D. and Kleinberg, J., The link prediction problem for social networks. In *Conference on Information and Knowledge Management*. ACM, 2003.

[19] Masry, E. and Tjøstheim, D., Nonparametric estimation and identification of nonlinear ARCH time series. *Econometric Theory*, 11:258–289, 1995. MR1341250

[20] Paparoditis, E. and Dimitris, N. P., The local bootstrap for markov processes. *J. Statist. Plann. Inference*, 108:301–328, 2002. MR1947405

[21] Pham, D., The mixing property of bilinear and generalised random coefficient autoregressive models. *Stochastic Processes and Their Applications*, 23:291–300, 1986. MR0876051

[22] Politis, D., Romano, J., and Wolf, M., *Subsampling*. Springer, 1999. MR1707286

[23] Politis, D. N. and Romano, J. P., Nonparametric resampling for homogeneous strong mixing random fields. *Journal of Multivariate Analysis*, 47(2):301–328, 1993. MR1247380

[24] Raftery, A. E., Handcock, M. S., and Hoff, P. D., Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 15:460, 2002. MR1951262

[25] Richard, E., Gaiffas, S., and Vayatis, N., Link prediction in graphs with autoregressive features. In P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2843–2851, 2012.

[26] Rinott, Y. and Rotar, V., A multivariate CLT for local dependence with $n^{-1/2}\log n$ rate and applications to multivariate graph related statistics. *Journal of Multivariate Analysis*, 56(2):333–350, 1996. MR1379533

[27] Sarkar, P., Chen, L., and Dubrawski, A., Dynamic network model for predicting occurrences of salmonella at food facilities. In *Biosurveillance and Biosecurity: International Workshop, BioSecure*. Springer, 2008.

[28] Sarkar, P. and Moore, A., Dynamic social network analysis using latent space models. In *Advances in Neural Information Processing Systems*. 2005.

[29] Snijders, T. and Nowicki, K., Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 1997. MR1449742

[30] Sunklodas, J., On normal approximation for strongly mixing random variables. *Acta Applicandae Mathematicae*, 97:251–260, 2007. MR2329733

[31] Tylenda, T., Angelova, R., and Bedathur, S., Towards time-aware link prediction in evolving social networks. In *ACM Workshop on Social Network Mining and Analysis*. ACM, 2009.

[32] Vu, D., Asuncion, A., Hunter, D., and Smyth, P., Continuous-time regression models for longitudinal networks. In *Advances in Neural Information Processing Systems*. MIT Press, 2011.

[33] Wang, M. C. and van Ryzin, J., A class of smooth estimators for discrete distributions. *Biometrika*, 1981. MR0614967

[34] Wilson, E., Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:209–212, 1927.

[35] Zhou, S., Lafferty, J., and Wasserman, L., Time varying undirected graphs. In *Conference on Learning Theory*, 2008.