AN INVESTIGATION OF SMOOTHNESS MEASURES FOR SEMI-SUPERVISED LEARNING ON GRAPHS

By

Ya Xu

Technical Report No. 2010-05
June 2010

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065
AN INVESTIGATION OF SMOOTHNESS MEASURES
FOR SEMI-SUPERVISED LEARNING ON GRAPHS

By

Ya Xu
Department of Statistics
Stanford University

Technical Report No. 2010-05
June 2010

This research was supported in part by
National Science Foundation grant DMS 0906056.

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu
AN INVESTIGATION OF SMOOTHNESS MEASURES FOR SEMI-SUPERVISED LEARNING ON GRAPHS

By Ya Xu *

Department of Statistics, Stanford University

In semi-supervised learning on graphs, response variables observed at one node are used to estimate missing values at other nodes. The methods depend on smoothness measures to regularize predictions toward smoother answers. In this paper, we investigate two popular choices of smoothness measures based on the graph Laplacian. We show on three graph datasets the surprising result that these measures can sometimes consider random response values to be much smoother than the ground truth. By applying Stein’s central limit theorem for dependent random variables, we are able to develop theoretical justifications for the empirical results, casting doubt on the utility of these measures.

1. Introduction. Over the past few years, data on graphs have grown tremendously in size and prevalence; consider, for instance, the explosion of social network data available on the Internet and gene interaction network data from Computational Biology. An important problem is to predict the value of a feature at one or more nodes in the graph. Usually, that feature has been measured on some, but not all nodes of the graph. For example, we might want to predict which web pages are spam, after a human expert has labeled a subset of them as spam or not. Similarly, we might want to know on which Facebook profile pages an ad would get a click, although that ad has only been shown on a subset of pages.

This problem is often called semi-supervised learning, because while the entire graph structure is available, the response values are only measured at some of the nodes. The underlying assumption in many prediction methods is that the response should vary smoothly in the graph. Such smoothness is usually achieved by regularization, and a popular choice is to use the graph Laplacian.

The smoothness assumption plays an essential role in the prediction algorithms. However, to the best of our knowledge, there has not been any systematic approach proposed to investigate whether a smoothness measure is

*Supported by NSF Grant DMS-0906056 and a Stanford Graduate Fellowship.
AMS 2000 subject classifications: 60F05, 68R10, 62H10, 62J20, 68U01
Keywords and phrases: graph Laplacian, smoothness measure, central limit theorem, semi-supervised learning, network
reasonable. Many graph-based learning algorithms attempt to validate their choices by demonstrating good empirical performance. Our approach starts by examining how much smoother the true labels are compared to random labels under a given measure. To our surprise, we found that randomizing the labels can sometimes lead to a smoother answer in real examples. This casts doubt on the utility of the corresponding smoothness measures, and motivated us to further investigate into the phenomenon.

The outline of this paper is as follows. Section 2 starts with the necessary notation and background on semi-supervised learning on graphs. In Section 3, we present the experimental results from three real datasets, where randomly permuted labels are sometimes found to be much smoother than the ground truth. We therefore propose a measure to quantify the validity of the smoothness assumption. Section 4 includes some theoretical results based on Stein’s central limit theorem for dependent random variables (Stein [1986]). It serves two purposes: first, it provides insight and theoretical justifications for the empirical results and our proposed measure; second, it shows that the surprising findings in Section 3 are reproducible.

2. Background. Let $G = (\mathcal{V}, \mathcal{E})$ be an undirected graph with $n$ nodes, where $\mathcal{V}$ and $\mathcal{E}$ denote the vertex set and the edge set. The graph $G$ is represented by a symmetric adjacency matrix $W$ with entries $w_{ij} > 0$ if there is an edge between $i$ and $j$, and $w_{ij} = 0$ otherwise. We impose $w_{ii} = 0$, so that if the graph contains loops, we do not count them. Node $i$ has degree $d_i = \sum_{j=1}^{n} w_{ij}$ and the volume of the graph is $\text{vol}(G) = \sum_{i=1}^{n} d_i$. We further assume that all $d_i > 0$ by removing any isolated nodes.

In practice, when the original data come in the form of a graph, the weight $w_{ij}$ usually has a natural interpretation. It could be the number of hyperlinks between two web pages, or a binary value indicating whether proteins $i$ and $j$ interact. Many times when the weights are not readily available from the data, they are computed based on symmetric and non-negative similarity measures. For instance, if each node lives in the Euclidean space $\mathbb{R}^d$, a popular choice is to use the Gaussian similarity measure $s_{ij} = \exp(-||x_i - x_j||^2/2\sigma^2)$, where $x_i \in \mathbb{R}^d$ describes the location of node $i$. The weights $w_{ij}$ are then constructed from $s_{ij}$ based on, for example, $\epsilon$-neighborhood or $k$-nearest neighbor.

Another representation of the graph $G$ is the graph Laplacian, which is also the main tool for many semi-supervised graph algorithms. For the purpose of this paper, we consider two definitions of graph Laplacian, the unnormalized
version $\Delta$ and the normalized version $\tilde{\Delta}$:

$$
\Delta \equiv D - W
$$

$$
\tilde{\Delta} \equiv D^{-1/2}(D - W)D^{-1/2} = I - D^{-1/2}WD^{-1/2},
$$

where $D = \text{diag}(d_1, \ldots, d_n)$. For properties of graph Laplacians see Luxburg [2007].

We suppose the feature of interest at node $i$ is $Y_i$. In many applications $Y_i \in \{-1, 1\}$ is binary, though we consider the general case of both discrete and continuous variables. In semi-supervised learning, we only observe realizations of $Y_i$ at some, but not all nodes of the graph. To make predictions on the unobserved features, the underlying assumption is that $Y_i$ and $Y_j$ are close in value if nodes $i$ and $j$ are close to each other in the graph. In other words, the vector $\mathbf{Y} \in \mathbb{R}^n$ is smooth with respect to a notion of distance in the graph. Mathematically, this smoothness assumption is usually reflected as a regularization term that penalizes large values of a variation functional.

Two popular choices for the functional correspond to the two versions of the graph Laplacian:

\[ \Omega(\mathbf{Y}) \equiv \mathbf{Y}^T \Delta \mathbf{Y} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (Y_i - Y_j)^2 \]

\[ \tilde{\Omega}(\mathbf{Y}) \equiv \mathbf{Y}^T \tilde{\Delta} \mathbf{Y} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left( \frac{Y_i}{\sqrt{d_i}} - \frac{Y_j}{\sqrt{d_j}} \right)^2. \]

Both smoothness measures penalize vectors $\mathbf{Y}$ that differ too much over similar nodes. However, by using the normalized graph Laplacian $\tilde{\Delta}$, $\tilde{\Omega}(\mathbf{Y})$ effectively scales the variables by $\sqrt{d_i}$.

For example, Belkin et al. [2004] defines the predictions to be

$$
\hat{\mathbf{Y}} = \arg\min_{\mathbf{Z} \in \mathbb{R}^n} \lambda \sum_{i \in \mathcal{L}} (Y_i - Z_i)^2 + \Omega(\mathbf{Z}),
$$

where $\mathcal{L}$ denotes the set of nodes with observed $Y_i$. The first term is to ensure the predictions to be close to the observations, and $\lambda > 0$ is a parameter that governs the trade-off between fit and smoothness. Other smoothness criteria have been used in the literature. Zhou et al. [2004] present a graph algorithm based on the normalized graph Laplacian and predict with

$$
\hat{\mathbf{Y}} = \arg\min_{\mathbf{Z} \in \mathbb{R}^n} \lambda \sum_{i=1}^{n} (Y_i - Z_i)^2 + \tilde{\Omega}(\mathbf{Z}).
$$

We do not intend to give a full review of semi-supervised algorithms. For a survey see Zhu [2005].
3. Motivation: Permutation experiment. Note the importance of the smoothness assumption in the semi-supervised learning algorithms. Through smoothing, the information from the observed nodes is “propagated” to the unobserved nodes. In fact, the smoothness measure such as the ones in (1) and (2) is usually the only component in the algorithms that connects the graph structure and the feature of interest. Clearly, if that measure fails, these algorithms are likely to perform poorly.

Moreover, the feature $Y$ may vary smoothly in the graph with respect to one smoothness measure, but less so to another. This is because there are many choices to make when defining a smoothness measure, and some may be better than others for a particularly dataset. First of all, there is no unique way to construct a graph when the edge weights $wij$ are not readily available. Readers are referred to a recent paper by Jebara et al. [2009] for a nice empirical study of several construction algorithms, while Argyriou et al. [2005] proposes to use an optimal combination of a number of differently constructed graphs. Secondly, the theoretical implications of Laplacian normalization are not well understood, thus it is unclear whether one should use the smoothness measure $\Omega(Y)$ or $\tilde{\Omega}(Y)$ in the learning algorithm. Johnson and Zhang [2007] derives near-optimal normalization factors under certain assumptions using a worst-case generalization bound.

The approach we take to investigate the smoothness measures starts with the following question: how much smoother are the true labels compared to a random set of labels? To be more concrete, for a fixed graph structure, we randomly permute the labels on the nodes and then compute the smoothness score. We then compare the score of the true labels with that of the permuted labels. This idea is demonstrated in Figure 1. Intuitively, permutation destroys the connection between the labels and the graph structure, and hence the resulted labels should be rougher with respect to the graph, resulting in a larger score.

As an attempt to validate our intuition, we perform this permutation procedure on three real datasets. The first dataset is a protein interaction network from Jeong et al. [2001], where some proteins are labeled as lethal and others are non-lethal. The second is a webspam dataset available at http://barcelona.research.yahoo.net/webspam/datasets/uk2007/. The third dataset is a snapshot of the Wikipedia website available at http://www.cise.ufl.edu/research/sparse/matrices/Gleich. This dataset originally does not include labels for each node. We construct labels as follows: (1) Set $Y_i = 1$ if node $i$ points to “United States” (the node with the highest indegree) and $-1$ otherwise. (2) Remove the “United States” node and all its edges. Table 1 includes a summary of all three datasets.
Table 1
Summary of the datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>Nodes</th>
<th>Edges</th>
<th>Response Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td>2114</td>
<td>4406</td>
<td>401 lethal (20.0%)</td>
</tr>
<tr>
<td>Webspam</td>
<td>3014</td>
<td>5578</td>
<td>225 spam (7.5%)</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>2,371,607</td>
<td>41,109,211</td>
<td>179,333 points to “US” (7.5%)</td>
</tr>
</tbody>
</table>

Figure 2 shows the results for the two smoothness measures defined in (1) (first row) and (2) (second row). The histograms represent the distributions of scores for the randomly permuted labels, with the red dashed lines being the medians of the empirical distributions. The red dots are the scores of the true labels. Under the smoothness assumption, we expect the ground truth to give smaller scores, however, five out of the six plots tell us otherwise. Except for plot (1a) where the red dot is in the left tail of the distribution, all the other red dots are either near the median line or deep in the right tail. This implies that the smoothness measures defined in (1) or (2) are sometimes inadequate to differentiate the true set of labels from a random permutation. Moreover, random shuffling could even result in labels that are much smoother as shown in (1c) and (2c)!

These results are quite striking. They not only reveal some potential problems with the smoothness assumption or some particular smoothness measures, but also inspire us to consider the following z-score to quantify how well the assumption holds:

$$z(\Omega, Y) = \frac{\Omega(Y) - \mathbb{E}(\Omega(Y^b))}{\text{sd}(\Omega(Y^b))},$$

where $Y^b$ is a bootstrap copy of the original response $Y$. The z-score quan-
tifies the difference between the red dot and the mean of the distribution, taking into account of the spread. Clearly, a large negative $z$-score supports the smoothness assumption, while a large positive one suggests otherwise.

In the next section, we will present some theoretical results that will justify the empirical findings in Figure 2 as well as the $z$-score defined in (3).

![Distributions of the smoothness scores of the randomly permuted labels, compared with that of the ground truth labels (red dots).](image)

**4. Central limit theorems for smoothness measures.** We notice that among the examples in Figure 2, almost all of the histograms look normally distributed. If we can well approximate the score distributions of the permuted labels with a normal distribution, the $z$-score defined in (3) is equivalent to a $p$-value in hypothesis testing, and therefore is a more interpretable measure that is readily comparable across different cases. However, the normal approximation clearly does not hold in general, as demonstrated in Figure 2 (2b). In this section, we will develop some theoretical results that give conditions under which normality holds.

Before we investigate how the smoothness measures behave asymptotically, we want to emphasize a characteristic that is common to almost all real graphs. Real graphs, particularly real large graphs, are usually very sparse, where the total number of edges in the graph grows linearly with $n$ (instead of $n^2$ as in a fully connected graph). We can notice this phenomenon
in the examples presented in Table 1. Another simple property of a graph is its degree sequence. Many studies have shown that real graphs tend to have a heavy tailed degree distribution (e.g. Albert et al. [1999]; Jeong et al. [2000]). Even though far from capturing the full structure of a graph, the degrees are fairly informative, and in fact, many efforts have been devoted to generating random graphs that match the degree distributions of those of real graphs.

To this end, we make two structural assumptions about the graph \( G \) in our analysis. First, we assume \( G \) is sparse such that \( \text{vol}(G) = \mathcal{O}(n) \). Second, the degrees of \( G \) have certain nice properties, and two different aspects are considered. In Section 4.2, we investigate the case where the maximal degree of \( G \) may grow with \( n \) while all other nodes have bounded degrees. We are able to give sufficient conditions where a normal approximation works or fails, which turn out to help explain our observations in Figure 2. In Section 4.3, the degrees are assumed to follow a probability distribution and the conditions for a CLT involve moments of that distribution. In the context of power-law degree distributions, these conditions can be readily translated into conditions on the exponent.

We simplify the analysis by assuming the graph is unweighted, i.e. \( w_{ij} \in \{0, 1\} \). All the results can be generalized to weighted graphs with nonzero weights bounded from above and away from zero. We further assume bounded feature values.

4.1. Preliminaries. We start with two preliminary theorems. Stein’s central limit theorem for dependent random variables is the most fundamental result of all those to be discussed in this section. It is quoted below:

**Theorem 4.1** (Stein [1986], p.110). Let \( X_1, \ldots, X_n \) be random variables, and let \( S_1, \ldots, S_n \) be subsets of \( \{1, \ldots, n\} \) such that

\[
\mathbb{E} X_i = 0, \quad \mathbb{E} X_i^4 < \infty, \quad \text{and} \quad \mathbb{E} \sum_{i=1}^n X_i \sum_{j \in S_i} X_j = 1.
\]
Let $T = \sum_{i=1}^{n} X_i$. Then for any $t \in \mathbb{R}$,

$$\left| \mathbb{P}(T \leq t) - \Phi(t) \right| \leq 2 \sqrt{\mathbb{E}\left\{ \sum_{i=1}^{n} \sum_{j \in S_i} (X_i X_j - \mathbb{E}X_i X_j) \right\}^2}$$

$$+ \sqrt{\frac{\pi}{2}} \mathbb{E} \sum_{i=1}^{n} \left| \mathbb{E}(X_i | X_j : j \notin S_i) \right|$$

$$+ 2^{3/4} \pi^{-1/4} \sqrt{\mathbb{E} \sum_{i=1}^{n} |X_i| \left( \sum_{j \in S_i} X_j \right)^2},$$

(4)

where $\Phi(\cdot)$ is the standard normal cdf.

Stein’s theorem allows dependence among the variables, and in many applications, $S_i$ contains all the variables $X_i$ depends on. Intuitively, the theorem states that if the overall dependence of the random variables is weak, a normal approximation of their sum still works well.

Rinott [1994] refines the result in Theorem 4.1 assuming that the random variables are bounded. Even though his intention was to improve the convergence rate in (4), we include this result here because it is directly applicable for the problems we consider in Section 4.2. We quote a necessary definition and then the theorem itself.

**Definition** Let $\{X_i : i \in V^*\}$ be a collection of random variables. The graph $G^* = (V^*, E^*)$, where $V^*$ and $E^*$ denote the vertex set and the edge set, respectively, is said to be a dependency graph for the collection if for any pair of disjoint subsets of $V^*$, $S_1$ and $S_2$ such that no edge in $E^*$ has one endpoint in $S_1$ and the other in $S_2$, the sets of random variables $\{X_i : i \in S_1\}$ and $\{X_i : i \in S_2\}$ are independent.

Note that we distinguish a dependency graph $G^*$ from a general graph $G$. $G^*$ is undirected and unweighted by definition. Moreover, it contains selfloops since a variable always depends on itself.

**Theorem 4.2** (Rinott [1994]). Let $X_1, \ldots, X_n$ be random variables having a dependency graph $G^*$ whose maximal degree is less than $d^*$, satisfying $|X_i - \mathbb{E}X_i| \leq c$ a.s. for $i = 1, \ldots, n$. Let $T = \sum_{i=1}^{n} X_i$ and assume $\text{var}(T) = \sigma^2 > 0$. Then

$$\left| \mathbb{P}\left( \frac{T - \mathbb{E}T}{\sigma} \leq t \right) - \Phi(t) \right| \leq \frac{1}{\sigma} \left( \sqrt{\frac{1}{2\pi}} d^* c + 16 \sqrt{\pi} d^{3/2} c^{2} + 10 \frac{n}{\sigma^2} d^{2} c^{3} \right).$$

(5)
Note that when $d^*$ and $c$ are bounded, the right hand side of (5) converges to zero at the rate of $n^{-1/2}$ if $\sigma^2 = \mathcal{O}(n)$.

4.2. Maximal degree conditions. As a starting point, we consider a simple case where all the degrees in the graph are bounded. Intuitively, a real person (i.e. not a robot) can only have so many real friends on Facebook even though the entire social network has millions of users. We then extend the results to a more general framework where the maximal degree is allowed to grow with $n$. This is a more realistic condition in many applications. For instance, on Internet graphs, there may be a few hub/authority websites that many other sites connect to, and it is clearly no longer appropriate to assume bounded degrees as $n \to \infty$.

We will first quote a result on variance of quadratic forms from Seber and Lee [2003]. We will find it useful when computing the variance of the smoothness scores.

**Lemma 4.3 (Seber and Lee [2003], p.10).** Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with mean $\mu_1$. Denote $\mu_r = \mathbb{E}(Y_i - \mu_1)^r$. If $A \in \mathbb{R}^{n \times n}$ is symmetric and $a = \text{diag}(A)$, then

$$\text{var}(Y^T A Y) = (\mu_4 - 3\mu_2^2)a^T a + 2\mu_3^2 \text{tr}(A^2) + 4\mu_2 \mu_1^2 1^T A^2 1 + 4\mu_3 \mu_1^2 1^T A a.$$

The following lemma states that the central limit theorem holds for both smoothness measures (1) and (2), when the maximal degree of a graph is bounded.

**Lemma 4.4 (Bounded maximal degree).** Let $Y_1, \ldots, Y_n$ be i.i.d. random variables corresponding to vertices of the graph $G$, satisfying $|Y_i| \leq c < \infty$ a.s. $\forall i$. Assume $\text{vol}(G) = \mathcal{O}(n)$ and $\max_i d_i \leq d < \infty$. Then, as $n \to \infty$

$$\frac{\Omega(Y) - \mathbb{E}\Omega(Y)}{\text{sd}(\Omega(Y))} \Rightarrow \mathcal{N}(0,1), \quad \text{and}$$

$$\frac{\tilde{\Omega}(Y) - \mathbb{E}\tilde{\Omega}(Y)}{\text{sd}(\tilde{\Omega}(Y))} \Rightarrow \mathcal{N}(0,1),$$

with $\Omega(\cdot)$ and $\tilde{\Omega}(\cdot)$ defined in (1) and (2), and $\mathcal{N}(0,1)$ being the standard normal distribution.

**Proof.** We prove the result in (6) and the similar proof for (7) can be found in Appendix A.1. Notice that $\Omega(Y)$ can be decomposed into sum of
dependent variables:

\[ \Omega(Y) = Y^T \Delta Y = \sum_{i=1}^{n} d_i Y_i^2 - \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} Y_i Y_j = \sum_{i=1}^{n} X_i, \]

where \( X_i = d_i Y_i^2 - \sum_{j=1}^{n} w_{ij} Y_i Y_j \). From here on, we will drop the summation limits when they go from 1 to \( n \).

The random variables \( X_1, \ldots, X_n \) are almost independent. In fact, \( X_i \) and \( X_j \) are dependent if and only if \( w_{ij} > 0 \) or \( \sum_k w_{ik} w_{kj} > 0 \). In other words, \( X_i \)'s have a dependency graph \( G^* \) where \( i \) and \( j \) are connected if and only if they are first or second order neighbors in the original graph \( G \). The maximal degree of this dependency graph is at most \( d + d^2 \), so is clearly bounded.

Therefore, (6) readily follows from Theorem 4.2 after we show that \( |X_i| \) is bounded a.s. for all \( i \) and \( \text{var}(\Omega(Y)) = O(n) \).

First,

\[ |X_i| \leq d_i Y_i^2 + |Y_i \sum_j w_{ij} Y_j| \leq 2c^2 d_i \leq 2c^2 d. \]

Applying Lemma 4.3 with \( A = \Delta \) and noting that \( \Delta_{ii} = d_i \) and \( \Delta 1 = 0 \), we have

\[ \text{var}(\Omega(Y)) = (\mu_4 - 3\mu_2^2) \sum_i d_i^2 + 2\mu_2 \sum_i \sum_j w_{ij}^2 \]

\[ = (\mu_4 - \mu_2^2) \sum_i d_i^2 + 2\mu_2 \sum_i \sum_j w_{ij}^2 \]

\[ \leq d(\mu_4 + \mu_2^2) \sum_i d_i \]

\[ = O(n), \]

where \( \mu_r \)'s are the centered moments of \( Y_i \) as defined in Lemma 4.3. \( \square \)

We are now ready to state a more general theorem where the maximal degree grows as \( O(n^{\epsilon}) \). It turns out that the asymptotic normality breaks down for \( \Omega(Y) \) when \( \epsilon \geq 1/2 \), while it holds for \( \tilde{\Omega}(Y) \) when \( 0 \leq \epsilon < 1 \). The theorem reduces to Lemma 4.4 when \( \epsilon = 0 \).

**Theorem 4.5.** Let \( Y_1, \ldots, Y_n \) be i.i.d. random variables corresponding to vertices of the graph \( G \), satisfying \( |Y_i| \leq c < \infty \) a.s. \( \forall i \). Assume \( \text{vol}(G) = O(n) \). Let \( d_k = \max_i a_i = O(n^{\epsilon}) \) with \( 0 \leq \epsilon < 1 \), while \( d_i \leq d < \infty \) \( \forall i \neq k \).
Then, as $n \to \infty$,

\[
\frac{\Omega(Y) - \mathbb{E}\Omega(Y)}{\text{sd}(\Omega(Y))} \implies \begin{cases} 
N(0, 1) & 0 \leq \epsilon < 1/2 \\
\mathcal{O}_p(1) \cdot N(0, 1) + \mathcal{O}_p(1) \cdot g(Y_k) & \epsilon = 1/2 \\
\mathcal{O}_p(1) \cdot g(Y_k) & 1/2 < \epsilon < 1,
\end{cases}
\]

where $g(Y_k) = Y_k^2 - 2\mu_1 Y_k - \mu_2^2 + \mu_1^2$.

\[
\frac{\tilde{\Omega}(Y) - \mathbb{E}\tilde{\Omega}(Y)}{\text{sd}(\tilde{\Omega}(Y))} \implies N(0, 1) \quad \forall \ 0 \leq \epsilon < 1,
\]

Proof. We prove the result in (10) and the proof for (11) can be found in Appendix A.2. Following equation (9), we have $\text{var}(\Omega(Y)) = \mathcal{O}(\sum_i d_i^2) = \mathcal{O}(n + n^2 \epsilon)$, where the last equality follows because $d_k = \mathcal{O}(n^\epsilon)$ is the only unbounded degree.

We can separate out the terms involving node $k$:

\[
\Omega(Y) = \Omega_{-k}(Y_{-k}) + \sum_i w_{ki}(Y_k - Y_i)^2,
\]

where $\Omega_{-k}(Y_{-k}) = \sum_{i \neq k} \sum_{j \neq k} w_{ij}(Y_i - Y_j)^2 / 2$. From here on, we will drop the subscript on $\Omega$ for clarity. Applying this trick to the standardized score, we get

\[
\frac{\Omega(Y) - \mathbb{E}\Omega(Y)}{\sqrt{n + n^2 \epsilon}} = \frac{\Omega(Y_{-k}) - \mathbb{E}\Omega(Y_{-k})}{\sqrt{n + n^2 \epsilon}} + \frac{\sum_i w_{ki} \left( (Y_k - Y_i)^2 - \mathbb{E}(Y_k - Y_i)^2 \right)}{\sqrt{n + n^2 \epsilon}}.
\]

By Lemma 4.4 the first term in (13) converges to a normal distribution when $\epsilon \leq 1/2$ and is $o_p(1)$ when $\epsilon > 1/2$. On the other hand,

\[
\frac{1}{\sqrt{n + n^2 \epsilon}} \sum_i w_{ki}(Y_k - Y_i)^2 = \frac{1}{\sqrt{n + n^2 \epsilon}} \sum_i w_{ki}(Y_k^2 - 2Y_k Y_i + Y_i^2)
\]

\[
= \frac{1}{\sqrt{n + n^2 \epsilon}} \left( d_k Y_k^2 - 2Y_k \sum_i w_{ki} Y_i + \sum_i w_{ki} Y_i^2 \right)
\]

\[
\implies \begin{cases} 
\mathcal{O}_p(1) & 0 \leq \epsilon < 1/2 \\
Y_k^2 - 2\mu_1 Y_k + (\mu_2 + \mu_1^2) & 1/2 \leq \epsilon < 1,
\end{cases}
\]

by Slutsky’s Theorem. The second term in (13) is this sum centered by its expectation, and hence goes to zero if $\epsilon < 1/2$ and to

\[
g(Y_k) = Y_k^2 - 2\mu_1 Y_k - (\mu_2 - \mu_1^2)
\]
when $\epsilon \geq 1/2$. Combining the two parts, together with $\text{var}(\Omega(Y)) = O(n + n^{2\epsilon})$, we get the desired result.

We can show that, by repeatedly applying (12), Theorem 4.5 generalizes to the case where there are a bounded number of nodes with degrees growing as $O(n^\epsilon)$.

It also follows from Theorem 4.5 that the smoothness measure $\tilde{\Omega}(Y)$ tends to be more robust than $\Omega(Y)$ against the maximal degree in terms of asymptotic normality. This is in fact well supported by the observations in Figure 2. The only non-normality occurs when applying $\Omega(Y)$ to the webspam dataset. A closer look at the webspam data shows that the maximal degree node is connected to about 25% of the entire graph. This is probably an artifact from web crawling, but nonetheless causes the maximal degree to be non-negligible compared to $n$. Even the two humps are well explained by the theorem. This is because for binary $Y_k$, the asymptotic distribution in (10) when $\epsilon = 1/2$ turns out to be a two-component Gaussian mixture.

4.3. Degree distribution conditions. Other than focusing on the maximal degree, we now consider the entire degree sequence. This is mainly motivated by the important role that degree distribution plays in the field of random graphs. The classical Erdős-Rényi model starts with $n$ nodes and connects every pair of nodes with probability $p$, producing degrees that follow a Poisson distribution. Recently, direct measurement of the degrees of many real graphs show that a Poisson distribution does not usually apply. Rather, the degrees tend to follow a power-law distribution. This finding then inspired the study of scale-free graph models that lead to a power-law degree distribution. Nevertheless, when graphs are characterized by their degree distributions, the results in Section 4.2 developed around the maximal degree are no longer sufficient. In the main theorem (Theorem 4.7) of this section, we will give conditions for a normal approximation that involve only the moments of the degree distribution.

For clarity and continuity, we include several supporting lemmas in Appendix A.3. The following lemma is an application of Stein’s result in Theorem 4.1, and is directly relevant to the main theorem to be introduced. It applies Stein’s theorem to the situation where the variables form a dependency graph with a prescribed degree sequence. Some of the proof techniques are borrowed from Baldi and Rinott [1989].

**Lemma 4.6.** Let $\{X_i, i \in V^*\}$ be random variables having a dependency graph $G^*$ with degrees $\{d_1^*, d_2^*, \ldots, d_n^*\}$. Let $T = \sum_i X_i$ and $S_i^* = \{j \in V^* : j \sim i\}$ be the set of nodes connected to node $i$. Assume that $\mathbb{E}X_i = 0$ and
\( \mathbb{E}(\sum_i \sum_{j \in S_i^*} X_iX_j) = 1 \). Then for any \( t \in \mathbb{R} \),

\[
\left| \mathbb{P}(T \leq t) - \Phi(t) \right| \leq 2^{3/4} \pi^{-1/4} \sqrt{\frac{1}{3} \sum_i \mathbb{E}|X_i|^3 d_i^2 + \frac{2}{3} \sum_i \sum_{j \in S_i^*} \mathbb{E}|X_i|^3 d_j^2}
\]

(14)

\[
+ 2 \sqrt{\sum_i \mathbb{E}X_i^4 \left( \sum_{j \in S_i^*} \sum_{k \in S_j^*} d_k^* + \sum_{k \in S_i^*} d_i^* d_k^* \right)}
\]

**Proof.** Following Theorem 4.1, it suffices to show that the RHS of (4) is bounded from above by the RHS of (14).

First,

\[
\mathbb{E} \sum_i |X_i| (\sum_{j \in S_i^*} X_j)^2 \leq \sum_i \sum_{j,k \in S_i^*} \mathbb{E}|X_i||X_j||X_k|
\]

\[
\leq \frac{1}{3} \sum_i \sum_{j,k \in S_i^*} \mathbb{E} (|X_i|^3 + |X_j|^3 + |X_k|^3)
\]

\[
= \frac{1}{3} \sum_i \mathbb{E}|X_i|^3 d_i^2 + \frac{2}{3} \sum_i \sum_{j \in S_i^*} \mathbb{E}|X_i|^3 d_j^2,
\]

where the inequality follows from (20) and the equality follows from equations (22) and (21) in the Appendix. On the other hand,

\[
\mathbb{E} \left( \sum_i \sum_{j \in S_i^*} (X_iX_j - \mathbb{E}X_iX_j) \right)^2 = \mathbb{E} \left( \sum_i \sum_{j \in S_i^*} X_iX_j \right)^2 - 1
\]

\[
\leq 2 \sum_i \mathbb{E}X_i^4 \left( \sum_{j \in S_i^*} \sum_{k \in S_j^*} d_k^* + \sum_{k \in S_i^*} d_i^* d_k^* \right),
\]

where the last step follows from (25) in the Appendix.

Finally, since \( X_i \) is independent of \( \{X_j : j \notin S_i^*\} \), the second term on the RHS of (4) vanishes. \( \Box \)

The following theorem gives conditions on the degree distribution in order to achieve normality. These conditions may seem stringent at first. However, we should keep in mind that the assumptions are on the degrees only, which is one of the simplest properties of a graph. The theorem needs to take into consideration the “worst” graph for a given degree sequence. It is possible to achieve sharper results by assuming more about the link structure of the graph.
Theorem 4.7. Let \( \{Y_i, i \in V\} \) be i.i.d. random variables on graph \( G = (V, E) \), whose degrees are \( \{d_1, \ldots, d_n\} \). Assume \( |Y_i| \leq c < \infty \) a.s. \( \forall i \), \( \text{vol}(G) = O(n) \) and \( \sum_i d_i^2 = O(n) \). As \( n \to \infty \),

(a) If \( n^{-3/2} \sum_i d_i^7 \to 0 \) and \( n^{-2} \sum_i d_i^{10} \to 0 \), then

\[
\frac{\Omega(Y) - \mathbb{E}\Omega(Y)}{\text{sd}(\Omega(Y))} \Rightarrow N(0, 1);
\]

(b) If \( n^{-3/2} \sum_i d_i^{11/2} \to 0 \) and \( n^{-2} \sum_i d_i^8 \to 0 \), then

\[
\frac{\tilde{\Omega}(Y) - \mathbb{E}\tilde{\Omega}(Y)}{\text{sd}(\tilde{\Omega}(Y))} \Rightarrow N(0, 1).
\]

Proof. We prove the result in (a) and the similar proof for (b) can be found in Appendix A.4.

Following the proof of Lemma 4.4, we write \( \Omega(Y) = \sum_i X_i \), where \( |X_i| \leq 2c^2 d_i \) and \( \sigma^2 \equiv \text{var}(\Omega(Y)) = O(\sum_i d_i^2) \) as shown in (8) and (9). The dependency graph formed by \( X_i \) has degrees \( d^*_i = d_i + \sum_{j \in S_i} d_j \), where \( S_i = \{j \in V : j \sim i\} \). We are now ready to apply Lemma 4.6 with the centered and scaled variables \((X_i - \mathbb{E}X_i)/\sigma\).

First,

\[
\sum_i \sigma^{-3} \mathbb{E}|X_i - \mathbb{E}X_i|^3 d_i^2 \leq \sigma^{-3}(4c^2)^3 \sum_i d_i^3 (d_i + \sum_{j \in S_i} d_j)^2
\]

\[
= 64c^6 \sigma^{-3} \sum_i \left( d_i^5 + 2 \sum_{j \in S_i} d_i^3 d_j + \sum_{j \in S_i} \sum_{k \in S_j} d_i^3 d_j d_k \right)
\]

\[
= O(\sigma^{-3} \sum_i d_i^7),
\]

where the last step follows from lemmas in A.1 and A.2. Similarly,

\[
\sum_i \sum_{j \in S_i} \sigma^{-3} \mathbb{E}|X_i - \mathbb{E}X_i|^3 d_j^3 = 64c^6 \sigma^{-3} \sum_i \left( \sum_{j \in S_i} d_j^3 + \sum_{j \in S_i} \sum_{k \in S_j} d_j^3 d_k \right)
\]

\[
= O(\sigma^{-3} \sum_i d_i^6),
\]
and
\[
\sum_{i} \sigma^{-4} \mathbb{E} |X_i - \mathbb{E}X_i|^4 \left( \sum_{j \in S_i^* k \in S_j^*} d_k^i + \sum_{k \in S_i^*} d_k^i \right)
\]
\[
\leq \sigma^{-4} (4c^2)^4 \sum_{i} d_i \left( \sum_{j \in S_i^* k \in S_j^*} d_k^i + \sum_{k \in S_i^*} d_k^i \right)
\]
\[
= \sigma^{-4} \mathcal{O} \left( \sum_{i} d_i \sum_{j \in S_i^*} \sum_{k \in S_j^*} d_k^i \right)
\]
\[
= \sigma^{-4} \mathcal{O} \left( \sum_{i} d_i \sum_{l \in S_i^* m \in S_j^* r \in S_k^*} \sum_{j \in S_j^*} \sum_{m \in S_m} \sum_{r \in S_r} d_{l,m} \right)
\]
\[
= \mathcal{O} (\sigma^{-4} \sum_{i} d_i^{10}),
\]
where the last step follows by applying (20) from Appendix repeatedly.

Putting together the three pieces and the assumption that \( \sigma^2 = \mathcal{O}(n) \), we have shown that the upper bound in (14) goes to 0, and hence the result in (a) follows.

Since the Erdős-Rényi random graphs have Poisson degree distributions with finite moments, the conditions in Theorem 4.7 are always satisfied. Unfortunately, this is usually not the case when the degrees follow the power-law
\[
\mathbb{P}(x) \propto x^{-\gamma}.
\]
Clearly, the asymptotic normality holds if \( \gamma > 11 \) or \( \gamma > 9 \) for (a) and (b) respectively. However, real graphs tend to have a smaller \( \gamma \) as shown in Bornholdt and Schuster [2003]. The following corollary obtains a better condition on \( \gamma \) using Theorem 8.9 in Durrett [2004] page 66.

**Corollary 4.8.** If \( \{d_i, i \in \mathcal{V}\} \) are i.i.d. and follow the power-law distribution in (15), then the conditions in Theorem 4.7 are equivalent to

(a') \( \gamma > 6 \),

(b') \( \gamma > 5 \).

**Proof.** Denote \( d_{i,m} = d_i^{m} \) for \( m > 1 \). Then
\[
\mathbb{P}(d_{i,m} = x) \propto x^{-\frac{m+\gamma-1}{m}}.
\]
Clearly, $E(d_{i,m})$ is finite if and only if $\gamma > m + 1$. When $E(d_{i,m}) = \infty$, it follows from Theorem 8.9 in Durrett [2004] that $\lim_{n \to \infty} n^{-q} \sum_{i=1}^{n} d_{i,m} = 0$ if and only if $\sum_{k=1}^{\infty} P(d_{i,m} \geq k^q) < \infty$ for $q > 1$. We can compute

$$\sum_{k=1}^{\infty} P(d_{i,m} \geq k^q) \propto \sum_{k=1}^{\infty} k^q \frac{(1-\gamma)}{m}.$$ 

It is well known that the sum of this infinite series converges if and only if $q(1 - \gamma)/m < -1$, i.e. $\gamma > 1 + m/q$. Plugging in the appropriate $m$ and $q$, we get the desired result.

5. Conclusion. Many recently developed semi-supervised learning methods for data on graphs assume that the feature of interest varies smoothly along the edges of the graph, and such smoothness has been used to motivate the Laplacian based regularization. Surprisingly, we have found that by randomly permuting the true labels on the graph, we can consistently get smoother answers in some real datasets. These examples strongly suggest inadequacy of the smoothness measures in many situations. To this end, we proposed the $z$-score in (3) to quantify how well the smoothness assumption holds. In an attempt to provide theoretical justifications for our proposal and also for the empirical observations, we have shown that both the unnormalized and normalized smoothness measures obey a central limit theorem under certain conditions on the node degrees.

If it is shocking that randomizing labels can give smoother results, then it is equally unsettling that how well the smoothness assumption stands is not indicative of performance. In some further empirical experiments, we compared the two smoothness measures from the following perspectives: the $z$-scores and the prediction errors. To our surprise, we found that the measure with a better $z$-score does not always perform better in terms of prediction. We have not yet been able to find a satisfactory explanation. We finish by pointing out that it is important to understand the smoothness properties of the data and to start with an assumption that truly reflects these properties.

Acknowledgments. I wish to thank Art B. Owen for many helpful and constructive suggestions.

APPENDIX

Proof. The proof follows similarly as that for (6). We first decompose \( \tilde{\Omega}(\mathbf{Y}) \) into a sum of dependent variables:

\[
\tilde{\Omega}(\mathbf{Y}) = \mathbf{Y}^T\tilde{\Delta}\mathbf{Y} = \sum_i Y_i^2 - \sum_i \sum_j \frac{w_{ij}}{\sqrt{d_id_j}} Y_i Y_j = \sum_i \tilde{X}_i,
\]

where \( \tilde{X}_i = Y_i^2 - \sum_j w_{ij}/\sqrt{d_id_j} Y_i Y_j \). Note that \( \tilde{X}_i \)'s have the same dependency graph as the \( X_i \)'s defined for \( \Omega(\mathbf{Y}) \). Again, the result readily follows from Theorem 4.2 after we show that \( |\tilde{X}_i| \) is bounded a.s. \( \forall i \) and \( \text{var}(\Omega(\mathbf{Y})) = \mathcal{O}(n) \).

Notice that

\[
\sum_j \frac{w_{ij}}{\sqrt{d_id_j}} = \frac{1}{\sqrt{d_i}} \sum_j \frac{w_{ij}}{\sqrt{d_j}} \leq \frac{1}{\sqrt{d_i}} d_i = \sqrt{d_i}.
\]

Therefore

\[
|\tilde{X}_i| \leq |Y_i^2| + \left| Y_i \frac{\sum_j w_{ij}}{\sqrt{d_id_j}} Y_j \right| \leq (\sqrt{d_i} + 1)c^2.
\]

Applying Lemma 4.3 with \( A = \tilde{\Delta} \) and noting that \( a = 1 \), we have

\[
\text{var}(\tilde{\Omega}(\mathbf{Y})) = (\mu_4 - 3\mu_2^2) n + 2\mu_2^2 \left( n + \sum_i \sum_j \frac{w_{ij}^2}{d_id_j} \right)
\]

\[
+ 4\mu_2\mu_1^2 \sum_i \left( 1 - \sum_j \frac{w_{ij}}{\sqrt{d_id_j}} \right)^2 + 4\mu_3\mu_1 \sum_i \left( 1 - \sum_j \frac{w_{ij}}{\sqrt{d_id_j}} \right)
\]

\[
= (\mu_4 - \mu_2^2 + 4\mu_2\mu_1^2 + 4\mu_3\mu_1) n + 2\mu_2^2 \sum_i \sum_j \frac{w_{ij}^2}{d_id_j}
\]

\[
+ 4\mu_2\mu_1^2 \sum_i \left( \sum_j \frac{w_{ij}}{\sqrt{d_id_j}} \right)^2 - (8\mu_2\mu_1^2 + 4\mu_3\mu_1) \sum_i \sum_j \frac{w_{ij}}{d_id_j}
\]

\[
= \mathcal{O}(n).
\]

The last equality follows from (16) and

\[
\sum_i \sum_j \frac{w_{ij}^2}{d_id_j} = \sum_i \frac{1}{d_i} \sum_j \frac{w_{ij}^2}{d_j} \leq \sum_i \frac{1}{d_i} d_i = n.
\]

Proof. The proof follows similarly as that for (10).
From (17), \( \text{var}(\hat{\Omega}(Y)) = O(\sum_i d_i) = O(n + n^2) = O(n) \). We separate out the terms involving node \( k \),

\[
\frac{\hat{\Omega}(Y) - E\hat{\Omega}(Y)}{\sqrt{n}} = \frac{\hat{\Omega}(Y_{-k}) - E\hat{\Omega}(Y_{-k})}{\sqrt{n}} + \frac{1}{\sqrt{n}} \sum_i w_{ki} \left( \frac{Y_k}{\sqrt{d_k}} - \frac{Y_i}{\sqrt{d_i}} \right)^2 - \frac{E(\frac{Y_k}{\sqrt{d_k}} - \frac{Y_i}{\sqrt{d_i}})^2}{\sqrt{n}}.
\]

(18)

By Lemma 4.4 the first term converges to a normal distribution. We now show that the second term in (18) is \( o_p(1) \). First,

\[
\frac{1}{\sqrt{n}} \sum_i w_{ki} (\frac{Y_k}{\sqrt{d_k}} - \frac{Y_i}{\sqrt{d_i}})^2 = \frac{1}{\sqrt{n}} \sum_i w_{ki} (Y_k^2/d_k - 2Y_i Y_k/\sqrt{d_k d_i} + Y_i^2/d_i)
\]

(19)

After centering, the third term in (19) becomes

\[
\frac{1}{\sqrt{n}} \sum_i w_{ki} (Y_i^2 - EY_i^2) = \frac{1}{\sqrt{s_k}} \sum_i w_{ki} (Y_i^2 - EY_i^2) = o_p(1).
\]

where \( s_k = \sum_i w_{ki}^2 / d_i^2 \leq d_k \). The last equality follows because \( s_k / n \to 0 \) and

\[
\frac{1}{\sqrt{s_k}} \sum_i w_{ki} (Y_i^2 - EY_i^2) \Rightarrow N(0, \text{var}(Y_i^2))
\]

by Billingsley [1995], Theorem 27.2. The second term in (19) goes to 0 because

\[
\frac{Y_k}{\sqrt{n} \sqrt{d_k}} \sum_i w_{ki} \frac{Y_i}{\sqrt{d_i}} = \sqrt{d_k} Y_k \left( \frac{1}{\sqrt{n}} \sum_i w_{ki} \frac{Y_i}{\sqrt{d_i}} \right) = \frac{\sqrt{d_k} Y_k}{\sqrt{n}} \cdot O_p(1) = o_p(1).
\]

The first term in (19) goes to 0 trivially because \( Y_i \) is bounded.
Collecting the three parts, we have proved that the second term in (18) is \( o_p(1) \), and hence the desired result. \( \Box \)
A.3. Supporting lemmas for Section 4.3.

**Lemma A.1.** For \( x_1, \ldots, x_m > 0 \), \( a_1, \ldots, a_m > 0 \) and \( a = \sum_{i=1}^{m} a_i \),

\[
\prod_{i=1}^{m} x_i^{a_i} \leq \sum_{i=1}^{m} \frac{a_i}{a} x_i^{a_i}.
\]

(20)

**Proof.** Applying Jensen’s inequality with the concave function \( \log(\cdot) \), we have

\[
\log\left(\sum_{i=1}^{m} \frac{a_i}{a} x_i^{a_i}\right) \geq \sum_{i=1}^{m} \frac{a_i}{a} \log(x_i^{a_i}) = \sum_{i=1}^{m} \log(x_i^{a_i}) = \log\left(\prod_{i=1}^{m} x_i^{a_i}\right).
\]

\[\square\]

**Lemma A.2.** Let \( G = (\mathcal{V}, \mathcal{E}) \) be a graph with degree sequence \( \{d_1, d_2, \ldots, d_n\} \). Let \( S_i = \{j \in \mathcal{V} : j \sim i\} \) be the set of nodes connected to node \( i \). Then,

\[
\sum_{j \in S_i} 1 = d_i,
\]

(21)

\[
\sum_{j \in S_i} \sum_{k \in S_i} 1 = d_i^2,
\]

(22)

\[
\sum_{i \in S_i} \sum_{j \in S_i} d_i d_j \leq \sum_{i} d_i^3,
\]

(23)

\[
\sum_{i \in S_i} \sum_{j \in S_i} \sum_{k \in S_j} d_k = \sum_{i \in S_i} \sum_{j \in S_i} d_i d_j \leq \sum_{i} d_i^3.
\]

(24)

**Proof.** The first two equalities are trivial. We now show (23) holds by applying (20) and then by symmetry,

\[
\sum_{i} \sum_{j \in S_i} d_i d_j \leq \frac{1}{2} \sum_{i} \sum_{j \in S_i} d_i^2 + d_j^2 = \sum_{i} \sum_{j \in S_i} d_i^2 = \sum_{i} d_i^3.
\]

We show the first half of (24), and the second half follows from (23).

\[
\sum_{i} \sum_{j \in S_i} \sum_{k \in S_j} d_k = \sum_{j \in \mathcal{V}: j \in S_i} \sum_{k \in S_j} d_k
\]

\[=\sum_{j \in \mathcal{V}: j \in S_i} \sum_{k \in S_j} d_k
\]

\[=\sum_{j \in \mathcal{V}: k \in S_j} \sum_{i \in S_i} 1
\]

\[=\sum_{j \in \mathcal{V}: k \in S_j} d_k d_j.
\]
**Lemma A.3.** Let \( \{X_i, i \in V^*\} \) be random variables having a dependency graph \( G^* \) with degrees \( \{d_1^*, d_2^*, \ldots, d_n^*\} \). Let \( S_i^* = \{j \in V^* : j \sim i\} \) be the set of nodes connected to node \( i \). Assume \( \mathbb{E}X_i = 0 \) and \( \mathbb{E}(\sum_i \sum_{j \in S_i^*} X_i X_j) = 1 \). Then,

\[
(25) \quad \mathbb{E}\left(\sum_i \sum_{j \in S_i^*} X_i X_j\right)^2 \leq 1 + 2 \sum_i \mathbb{E}X_i^4 \left(\sum_{j \in S_i^*} \sum_k d_k^* + \sum_{k \in S_i^*} d_i^* d_k^*\right)
\]

**Proof.** First,

\[
\mathbb{E}\left(\sum_i \sum_{j \in S_i^*} X_i X_j\right)^2 = \sum_{\{i,j\}\{k,l\}} X_i X_j X_k X_l + \sum_{\{i,j,k,l\}} X_i X_j X_k X_l,
\]

where \( j \in S_i^* \) and \( l \in S_k^* \) in both sums, but \( \{i,j\} \) and \( \{k,l\} \) form a connected set whereas \( \{i,j\} \) and \( \{k,l\} \) are disconnected. Notice that

\[
\mathbb{E}\sum_{\{i,j\}\{k,l\}} X_i X_j X_k X_l = \mathbb{E}\left(\sum_i \sum_{j \in S_i^*} X_i X_j\right)\mathbb{E}\left(\sum_k \sum_{l \in S_k^*} X_k X_l\right) = 1,
\]

and

\[
\mathbb{E}\sum_{\{i,j,k,l\}} X_i X_j X_k X_l \leq \frac{1}{4} \mathbb{E}\left(\sum_i \sum_{j \in S_i^*} X_i^4 + X_j^4 + X_k^4 + X_l^4\right) = \sum_{\{i,j,k,l\}} \mathbb{E}X_i^4,
\]

where the last step follows by symmetry.

For a fixed node \( i \), \( \{i,j\} \) and \( \{k,l\} \) are connected only if the four nodes have at least one of the following connections: \( i \sim j \sim k \sim l \) or \( j \sim i \sim k \sim l \), where \( k \) and \( l \) are interchangeable. This notation includes the case where the four tuple includes only 2 or 3 distinct nodes. Therefore,

\[
\sum_{\{i,j,k,l\}} \mathbb{E}X_i^4 \leq 2 \sum_i \mathbb{E}X_i^4 \left(\sum_{j \in S_i^*} \sum_k d_k^* + \sum_{k \in S_i^*} d_i^* d_k^*\right)
\]

Collecting the two pieces we get (25).  \( \square \)

Proof. Following the proof of Lemma 4.4, we write \( \tilde{\Omega}(\mathbf{Y}) = \sum_i \tilde{X}_i \), where \( |\tilde{X}_i| \leq (\sqrt{d_i} + 1)c^2 \leq 2c^2\sqrt{d_i} \) and \( \sigma^2 \equiv \var(\tilde{\Omega}(\mathbf{Y})) = \mathcal{O}(n) \). The dependency graph formed by \( \tilde{X}_i \) has degrees \( d^*_i = d_i + \sum_{j \in S_i} d_j \). We are now ready to apply Lemma 4.6 with the centered and scaled variables \((\tilde{X}_i - \mathbb{E}\tilde{X}_i)/\sigma\).

First, 

\[
\sum_i \sigma^{-3} \mathbb{E} \left| \tilde{X}_i - \mathbb{E}\tilde{X}_i \right|^3 d_i^{1/2} \leq \sigma^{-3}(4c^2)^3 \sum_i d_i^{3/2}(d_i + \sum_{j \in S_i} d_j)^2 \\
= 64c^6 \sigma^{-3} \sum_i d_i^{3/2} \left( d_i^2 + 2 \sum_{j \in S_i} d_j d_j + \sum_{j \in S_i} \sum_{k \in S_i} d_j d_k \right) \\
= \mathcal{O}(\sigma^{-3} \sum_i d_i^{11/2}),
\]

where the last step follows from Lemma A.1 and A.2. Similarly,

\[
\sum_{i,j} \sigma^{-3} \mathbb{E} \left| \tilde{X}_i \tilde{X}_j \right|^3 d_j^{1/2} = 64c^6 \sigma^{-3} \sum_{i,j} \left( \sum_{j \in S_i} d_j^{3/2} + \sum_{j \in S_i} \sum_{k \in S_j} d_j d_k \right) \\
= \mathcal{O}(\sigma^{-3} \sum_i d_i^{9/2}),
\]

and

\[
\sum_i \sigma^{-4} \mathbb{E} |X_i - \mathbb{E}X_i|^4 \left( \sum_{j \in S_i^*} \sum_{k \in S_i^*} d_k^d + \sum_{k \in S_i^*} d_k^d d_k^d \right) \\
\leq \sigma^{-4}(4c^2)^4 \sum_i d_i^2 \left( \sum_{j \in S_i^*} \sum_{k \in S_i^*} d_k^d + \sum_{k \in S_i^*} d_k^d d_k^d \right) \\
= \sigma^{-4} \mathcal{O} \left( \sum_i d_i^2 \sum_{j \in S_i^*} \sum_{k \in S_i^*} d_k^d \right) \\
= \sigma^{-4} \mathcal{O} \left( \sum_i d_i^2 \sum_{j \in S_i^*} \sum_{m \in S_j} \sum_{k \in S_m} \sum_{r \in S_k} d_r \right) \\
= \mathcal{O}(\sigma^{-4} \sum_i d_i^8),
\]

where the last step follows by applying (20) repeatedly.

Putting together the three pieces and the assumption that \( \sigma^2 = \mathcal{O}(n) \), we have shown that the upper bound in (14) goes to 0, and hence the result in (b) follows. \(\square\)
REFERENCES


Department of Statistics
390 Serra Mall
Stanford University
Stanford, CA 94305-4065
E-mail: yax@stanford.edu