Spectral Analysis of Random Graphs with Skewed Degree Distributions

Anirban Dasgupta* John E. Hopcroft† Frank McSherry‡

Abstract

We extend spectral methods to random graphs with skewed degree distributions through a degree based normalization closely connected to the normalized Laplacian. The normalization is based on intuition drawn from perturbation theory of random matrices, and has the effect of boosting the expectation of the random adjacency matrix without increasing the variances of its entries, leading to better perturbation bounds.

The primary implication of this result lies in the realm of spectral analysis of random graphs with skewed degree distributions, such as the ubiquitous "power law graphs". Recently Mihail and Papadimitriou [22] argued that for randomly generated graphs satisfying a power law degree distribution, spectral analysis of the adjacency matrix will simply produce the neighborhoods of the high degree nodes as its eigenvectors, and thus miss any embedded structure. We present a generalization of their model, incorporating latent structure, and prove that after applying our transformation, spectral analysis succeeds in recovering the latent structure with high probability.

1. Introduction

The analysis of large networks has come to the forefront of much research, with examples such as the internet, social networks, bibliographic databases, energy distribution networks, and global networks of economies motivating the development of the field. The ubiquity of large networks in social and technological fields makes the analysis of the common properties of these networks important. A central problem in this setting is that of finding communities, clusters, and other latent structure.

Spectral analysis is a widely used heuristic for a range of clustering problems. However, there have been relatively few theoretical results to support its use (notable exceptions include: [30, 19]). As it appears difficult to make general claims about worst-case data, researchers have looked to analyzing spectral algorithms applied to generative models for graphs, specifically with an eye towards recovering planted latent structure [26, 6, 2]. In particular, Azar et al argue in [6] that reconstruction through spectral methods works well for graphs produced by independent random rounding of the entries of a low rank matrix. However, their results make significant assumptions about the magnitudes of elements in the matrix and their reconstructive result does not apply to the degree distributions found in many real networks, which tend to exhibit a heavy tail.

Indeed, the effects of heavy tailed distribution of data have long plagued practitioners of spectral analysis. Ding et al [11] discuss the effect of term-frequency distributions on latent semantic indexing and showed that the current techniques for normalizing term frequencies are often insufficient to remove a spectral bias towards the high frequency terms. A recent paper of Mihail and Papadimitriou [22] provides a more damning argument: In random graphs where the maximum degrees far outstrip the average degree, the largest eigenvalues are in correspondence with the nodes of highest degrees, and the corresponding eigenvectors are simply the characteristic vectors of their neighborhoods. In particular, their results apply to graphs with power-law distributed vertex degrees, as most social networks appear to be. While their results were argued for purely random graphs without any latent structure, they apply equally well to structured random graphs with skewed degree distribu-

* Dept. of Computer Science, email:adg@cs.cornell.edu. Supported by Department of Computer Science.

† Dept. of Computer Science,email:jeh@cs.cornell.edu. Work supported in part by NSF IIS-0121175, NSF IIS-0312910 and the Intelligent Information Systems Institute at Cornell University sponsored by AFOSR(F 49620-01-0076)

‡ Microsoft Research, email:mcsherry@microsoft.com

tions. On such graphs with skewed degrees, spectral analysis ignores the graph's latent structure in favor of high degree nodes. Needless to say, this caused some concern in the spectral analysis community, as it effectively implies that spectral methods will simply fail in these domains.

In this paper we develop a normalization technique based on intuition drawn from the perturbation theory of random matrices, and closely connected to the normalized laplacian in the case of undirected graphs. We demonstrate the efficiency of the normalization technique by applying it to a model for skewed degree structured random graphs based on the planted partition model[21]. We prove that the technique reconstructs the latent partition in the face of dramatically skewed degrees, and can even take advantage of the skewing to yield better bounds.

1.1. The Normalization

Rather than analyze a graph's adjacency matrix, we apply spectral analysis to an alternate matrix, constructed by reweighting the edges of the graph. Let $\widehat{M}$ be the input adjacency matrix[1] and D_r and , D_c be diagonal matrices whose entries are respectively the degrees of the rows and columns in $\widehat{M}$. The normalization we consider is the matrix

$$\widehat{L} = D_r^{-1/2} \widehat{M} D_c^{-1/2}. \tag{1}$$

Stated alternately, $\widehat{L}$ is constructed by dividing each entry of $\widehat{M}$ by the geometric mean of its corresponding row and column degrees.

The matrix $\widehat{L}$ is closely related to the normalized Laplacian of a graph, which for symmetric graphs is equal to $I - \widehat{L}$. We prefer the matrix $\widehat{L}$ for two significant reasons: First, the normalized laplacian does not extend well to directed and bipartite graphs, both extensively used in analyzing LSI [15], recommendation systems, and web search [6, 25, 20]. Second, the important spectral structure of $I - \widehat{L}$ is found in its eigenvectors of smallest eigenvalue. By considering $\widehat{L}$ instead, the eigenvectors are promoted to the most significant eigenvalues, and the techniques of matrix perturbation theory [10, 31] and that of near optimal low rank approximations [14, 2] translate effortlessly.

A natural motivation for using $\widehat{L}$ in place of $\widehat{M}$ is found in the results of Azar et al [6], who bound the perturbation of a random matrix from its expectation by the largest variance of any of its entries. With this bound in hand, whenever the variances of $\widehat{M}$ are not uniform one can imagine scaling the low variance entries up until all variances are of similar size. While this will not affect the bound on the perturbation, it does inflate the expected matrix against which the perturbation is measured. If one views the entries in a row u as the result of d_u samples from an arbitrary distribution, the variances of the $\{0, 1\}$ entries in row u are proportional to d_u. Multiplying an entry by $(d_{\max}/d_u)^{1/2}$ multiplies its variance by $d_{\max}/d_u$, and is therefore the correct factor to apply to each row to unify the variances. An analogous argument justifies applying $(d_{\max}/d_v)^{1/2}$ to each column v. Modulo the $d_{\max}^{1/2}$ factor, this is what our normalization accomplishes.

1.2. Related Work

Theoretical analysis of spectral algorithms applied to large random graphs is an active research area [26, 1, 6, 19]. An important issue is how to normalize the data before applying spectral techniques. Practitioners of information retrieval have addressed the issue of distribution of document lengths and term frequencies by various heuristics [33, 18]. These heuristics have had varying experimental success but there is a lack of provable bounds on their performance.

Ding et al [11] noted that the effect of term-frequencies on singular values often survives current normalization techniques. Mihail and Papadimitriou [22] observed that in random graphs with power law degree distributions, the degree sequence essentially determines the large eigenvalues, and the eigenvectors that result are of little value.

Use of the normalized Laplacian has been promoted by several practitioners of spectral analysis, [13, 12, 8]. The works of [13, 12] examine the performance of the normalized Laplacian in spectral algorithms for multi-partition. In these, as in other work examining the performance of spectral heuristics, the use of the normalized Laplacian has largely been inspired by the idea of low conductance graph bisection. Chung et al[8], studied the eigengap in the normalized Laplacian for the Mihail/Papadimitriou random graph model.

The SALSA algorithm of Lempel and Moran [27] for ranking web pages is very closely related to our approach. SALSA weights pages by conducting a backward-forward random walk by alternately following random in-links and out-links, ranking pages by their stationary probability. In [27, 28] it is argued that SALSA is the first algorithm that is stable on a certain class of authority based graphs (those graphs for which their random walk is connected), whereas HITS [20], for example, is not. As it turns out, their score converges to the principal right singular vector of $\widehat{L}$, scaled by a $D^{1/2}$ term, whereas HITS converges to the principal right singular vectors of $\widehat{M}$. This appears to further support the use of $\widehat{L}$.

[1] We use the $\widehat{\cdot}$ notation on capitalized letters to refer to random matrices. While this may appear cumbersome, we will make much use of expected matrices, represented by the unadorned corresponding letters.

1.3. The Model

We will briefly describe the random graph model of McSherry [21], called the “Planted Partition Model”, as this will be our starting point in establishing a graph model that supports skewed degree distributions. We note that there is a large community of researchers investigating models for skewed degree random graphs, but we will focus on the planted partition model as it lends itself to spectral analysis expressly.

The underlying assumption in the planted partition model is that each directed edge (u, v) is present independently with probability G_{uv}, and that the matrix of probabilities G is a block matrix that has k row and k column blocks. The adjacency matrix $\widehat{G}$ is then formed by randomly rounding the entries of G:

$$\widehat{G}_{uv} = \begin{cases} 1 & \text{with probability } G_{uv} \\ 0 & \text{otherwise} \end{cases}$$

Thus, there are latent partitions ψ_r and ψ_c of the m rows and n columns into k parts. The presence of each edge is independent events, whose probability depend only on the *parts* to which each of its indices belong.

It is worth noting that the degrees that we expect to see in this model are fairly uniform. For each part, each node has an identical distribution over edges, and since the minimum degrees allowed by the model of [21] is at least $(\log n)^6$, these distributions are concentrated.

Extended Planted Partition problem : We extend the planted partition model by specifying scalars $d_u \geq 1$ and $d_v \geq 1$ for each row u and column v corresponding to their intended degrees. We scale the probabilities in G by the values d_u and d_v and form the new matrix $\widehat{M}$ distributed as:

$$\widehat{M}_{uv} = \begin{cases} 1 & \text{with probability } d_u G_{uv} d_v \\ 0 & \text{otherwise} \end{cases}$$

Letting $M_{uv} = d_u G_{uv} d_v$ denote the probability that edge $\widehat{M}_{uv}$ exists, we can write $M = D_r G D_c$, where D_r and D_c are diagonal matrices of the $\{d_u\}$ and $\{d_v\}$ values.

In this formulation the expected degree of a row u is $d_u \sum_v G_{uv} d_v$. The sum $\sum_v G_{uv} d_v$ is the same for all rows in the part $\psi_r(u)$, so while the expected degrees of the nodes do not necessarily equal d_u, the degrees of nodes u_1 and u_2 within a part *are proportional* to d_{u_1} and d_{u_2}. It is to this extent that we represent arbitrary skewed degree distributions.

Notation: A word about notation is needed here. We refer to column u of a matrix X by X_u and row u of X by X_u^T. We let $d_{\max}$ and $d_{\min}$ be the greatest and least d_u, d_v of any row or column. We will let s_u be the total degree of the part to which u belongs, and $s_{\min}$ equal the least total degree of any part. For a set of columns A, we let $s^A_{\min}$ refer to the minimum total degree of parts in A. In the case of symmetric probability matrices, we refer to the partition of nodes as ψ. We will also use the notations d_u and s_u loosely, to refer to either or a row or column scalar depending on the use of u as a row or column index, which should be clear from context.

1.4. Statement of Results

The main result of the paper states that we can extract the partition ψ_r and ψ_c from an observed matrix $\widehat{M}$, provided the expected row and column vectors of different parts are separated in a normalized sense. We start with a theorem that addresses the case of symmetric M, and then follow with a corollary stating the result for non-symmetric M.

For reasons of analysis, we must restrict our theorem to those degree distributions $\{d_u\}, \{d_v\}$ with the property that a uniformly random bipartitioning of the rows or columns yields parts A, B for which $s^A_u > s_u/4$ and $s^B_u > s_u/4$ for all u, with probability at least $3/4$.

Theorem 1 *Let G, D, M come from an instance of the planted partition problem where $M = M^T$, and let $\sigma^2 \gg (\log n)^6/n$ be an upper bound on the entries in G. There is a constant c such that for sufficiently large n, if when $\psi(u) \neq \psi(v)$*

$$|D^{1/2}(G_u - G_v)|^2 \geq c\sigma^2 k \log k \left(\frac{n}{s_{\min}} + \frac{\log(n/\delta)}{d_{min}} \right)$$

then with probability $1 - \delta - n^{-\omega(1)}$, we can efficiently recover ψ from $\widehat{M}$, D, k, $s_{\min}$, and σ.

The special case of the uniform degree planted partition problem handled in [21] can be obtained by instantiating $d_u = 1$ for all u. Furthermore, as the d_u increase from 1, the difference on the left hand side becomes more pronounced and the required separation on the right hand side becomes smaller, demonstrating that increased d_u can actually help recover latent structure.

The non-symmetric form of the result is similar, only with more notation.

Corollary 2 *Let $G \in \mathbb{R}^{m \times n}, D, M$ come from an instance of the planted partition problem, and assume $n \geq m$. Let $\sigma^2 \gg (\log n)^6/n$ be an upper bound on the entries in G. There is a constant c such that for sufficiently large m, n, if when $\psi_r(u) \neq \psi_r(v)$*

$$|(G_u^T - G_v^T)D_c^{1/2}|^2 \quad \geq \quad c\sigma^2 k \log k \left(\frac{n}{s_{\min}} + \frac{\log(n/\delta)}{d_{min}} \right)$$

and when $\psi_c(u) \neq \psi_c(v)$

$$|D_r^{1/2}(G_u - G_v)|^2 \quad \geq \quad c\sigma^2 k \log k \left(\frac{n}{s_{\min}} + \frac{\log(n/\delta)}{d_{min}} \right)$$

then with probability $1 - \delta - n^{-\omega(1)}$*, we can efficiently recover* ψ_r *and* ψ_c *from* $\widehat{M}$*,* D_r*,*D_c*,* k*,* $s_{\min}$*, and* σ*.*

While the scalars d_u may not be known, knowledge of the expected degrees of the graph is sufficient. Letting $d_u = \sum_{v=1}^{n} M_{uv}$ be the expected degree of node u, defining d_{avg} as the average expected degree and $d_{\min}$ and $s_{\min}$ as above, we have the following corollary.

Corollary 3 *Let* $M \in \mathbb{R}^{n \times n}$ *be defined as in the extended planted partition problem, and let* d_u *be the expected degree of node* u*. If there is a constant* c *such that for sufficiently large* n*, when* $\psi(u) \neq \psi(v)$

$$|M_u/d_u - M_v/d_v|_1 \geq c\sqrt{kd_{avg}}\left(\frac{n}{s_{\min}} + \frac{\log^6(n/\delta)}{d_{\min}}\right)$$

then with probability $1 - \delta - n^{-\omega(1)}$*, we can efficiently recover* ψ *from* $\widehat{M}$*,* D*,* k*,* $s_{\min}$*, and* σ*.*

Notice that M_u/d_u is the probability distribution of the edges for node u. If each part represents a substantially different distribution, we will be able to distinguish one part from another and reconstruct ψ. In interpreting this result, notice that if the right hand side is more than 1, the condition is never satisfied. In order for the right hand side to be less than 1, it must be the case that $d_{\min} \gg \log^6 n$. This implies that our analysis does not apply to overly sparse graphs. Even when the graph is dense, we must concern ourselves with the relative sizes of $d_{\min}$, d_{avg}, and $s_{\min}$. The corollary derives from Theorem 1, and an analogous form can be derived from Corollary 2.

1.5. Paper Outline

In Section 2 we motivate and present our algorithm for extracting the column partition, followed by a proof of its correctness in Section 2.2. In Section 3 we discuss some extensions, and propose future research.

2. Algorithm and Proof

We start with a review of some linear algebra tools. An important notion in spectral analysis is that of an optimal rank k approximation, which for any matrix $\widehat{X}$ is the rank k matrix $\widehat{X}^{(k)}$ minimizing $\|\widehat{X} - \widehat{X}^{(k)}\|_F$. This matrix can also be described as the *projection* of $\widehat{X}$ onto the k-dimensional subspace spanned by the first k singular column vectors. That is,

$$\widehat{X}^{(k)} = P_{\widehat{X}}^{(k)}(\widehat{X})$$

where $P_{\widehat{X}}^{(k)}$ is the projection onto the the span of the first k singular column vectors. The reader can find more details in [6], or in the text of [17].

Our work builds on a result of McSherry [21], who observes that the optimal rank k approximation to a random matrix is a good approximation to the expected matrix, if that matrix has rank at most k. In [21] it is proven that

Theorem 4 *Let* $\widehat{X}$ *be a* $m \times n$ *matrix whose entries are independent random variables concentrated on a unit interval. Let* $\sigma^2 \gg \log^6(m+n)/(m+n)$ *be an upper bound on the variances of the entries of* $\widehat{X}$*. If* $X = E[\widehat{X}]$ *has rank at most* k*, then with probability at least* $1 - 2e^{-\sigma^2(m+n)/4}$*,*

$$\|X - \widehat{X}^{(k)}\|_F^2 \quad \leq \quad 128\,\sigma^2 k(m+n)\ .$$

This theorem bounds the difference between the computable matrix $\widehat{X}^{(k)}$, and the matrix of expectations X that reveals the latent rank k structure. McSherry shows in [21] that when the variances of $\widehat{G}$ are small enough compared to the row and column separation in the expected matrix G, the partitions ψ_r, ψ_c can be recovered from $\widehat{G}$.

For the skewed degree model, note that for $M_{uv} \ll 1$,

$$\sigma^2(\widehat{G}_{uv}) \approx G_{uv} \quad \text{but} \quad \sigma^2(\widehat{M}_{uv}) \approx d_u G_{uv} d_v\ \ .$$

The maximum variance of the $\widehat{M}_{uv}$, on which Theorem 4 is based, increases in proportion to $d_{\max}^2$, while the average of the entries, $M_{uv} = d_u G_{uv} d_v$, only increase in proportion to d_{avg}^2. For highly skewed degree distributions, the error can easily overwhelm the latent structure of the expected matrix, thereby causing the approach of [21] to fail.

With the bound of Theorem 4 in mind, we now motivate the normalization by showing that it undoes the skewing of variances brought on by the skewed d_u, d_v, and can actually improve performance. Recalling that $\widehat{L} = D_r^{-1/2}\widehat{M}D_c^{-1/2}$, observe that its entries $\widehat{L}_{uv}$ satisfy

$$\sigma^2(\widehat{L}_{uv})\ \leq\ \sigma^2(\widehat{G}_{uv}) \quad \text{and} \quad E[\widehat{L}_{uv}]\ \geq\ E[\widehat{G}_{uv}]\ ,$$

with equality only when $d_u = d_v = 1$. Comparing $\widehat{L}$ with $\widehat{G}$, the entries of $\widehat{L}$ have favorable variance and expectation. Moreover, $L = E[\widehat{L}]$ is a rank k matrix exhibiting the same latent structure as G, albeit scaled by $D_r^{1/2}$ and $D_c^{1/2}$. Intuitively, it should be as easy or easier to recover latent structure from $\widehat{L}$ than from $\widehat{G}$, and indeed we show this is the case in the planted partition setting.

For clarity and future reference, we note down the following corollary of Theorem 4.

Corollary 5 *Let the matrices* D_r*,* D_c*,* G *and* $M = D_r G D_c$ *be defined as in the* **extended planted partition** *problem, and let* $\sigma^2 \gg (\log n)^6/n$ *bound the largest entry in* G*. Defining*

$$\widehat{L} = D_r^{-1/2}\widehat{M}D_c^{-1/2} \quad \textit{and} \quad L = D_r^{-1/2}MD_c^{-1/2}\ ,$$

with probability at least $1 - 2e^{-\sigma^2 n/4}$*,*

$$\|L - \widehat{L}^{(k)}\|_F^2 \quad \leq \quad 256\,\sigma^2 kn\ .$$

It is important to note that this corollary represents more than just a slight tweaking of the parameter space for which the spectral approach applies. Graphs with heavy tailed degree distributions have maximum degree *much* larger than their average degree, a polynomial factor in n for power-law distributions. As is argued in [22], this difference is significant enough in many domains to render spectral techniques useless when applied to $\widehat{M}$.

2.1. Our Algorithm

In the following section we describe our algorithm and present an analysis of it. The algorithm and its analysis show how to extract the column partitions ψ_c, the case for row partitions being completely analogous.

The basic spirit of our algorithm is fairly simple: Given $\widehat{M}$, D_r, D_c we form the matrix $\widehat{L}$ and compute a low rank approximation to it. At this point, we can apply a greedy clustering algorithm to the columns of the projected matrix yielding ψ_c. For technical reasons, we will actually first partition the data set into two pieces and cross-train to avoid conditioning issues, arriving at the algorithm

```
Normalized Partition ($\widehat{M}$, $D_r$, $D_c$, $\tau$):
1. Let $\widehat{L} = D_r^{-1/2}\widehat{M}D_c^{-1/2}$ and $\widehat{N} = \widehat{L}D_c^{-1/2}$.
2. Randomly partition the columns of $\widehat{L}$ into $\widehat{L}_A, \widehat{L}_B$.
3. Compute
       $Q_A^{(k)} = \textbf{Projection}(D_r^{-1/2}\widehat{L}_A^{(k)}, \tau)$
       $Q_B^{(k)} = \textbf{Projection}(D_r^{-1/2}\widehat{L}_B^{(k)}, \tau)$
4. Compute a minimum spanning tree on the columns of
       $F = [Q_B^{(k)}\widehat{N}_A, Q_A^{(k)}\widehat{N}_B]$ .
   Sever the $k-1$ heaviest edges, and return the partition
   of columns defined by the connected components.
```

The method **Projection**($\widehat{X}, \tau$), for reasons of analysis, computes a column projection from a clustering the *rows* of $\widehat{X}$, in which the characteristic vector of each cluster serves as a basis vector. The clustering of rows is performed greedily, repeatedly selecting a row and extracting all of its nearby neighbors into a cluster.

```
Projection($\widehat{X}, \tau$):
1. Let all rows be initially unmarked.
2. For $i$ from 1 to $k$:
   (a) Let $u_i$ be a random unmarked row, choosing row
       $u$ with probability proportional to $d_u$.
   (b) Mark each row $v$ for which $|\widehat{X}_v^T - \widehat{X}_{u_i}^T| < \tau$.
3. Define $\widehat{\psi}_r$ by assigning each row $v$ to its closest center.
   Specifically, set $\widehat{\psi}_r(v) = argmin_i\, |\widehat{X}_v^T - \widehat{X}_{u_i}^T|$.
4. Let $\widehat{y}_i$ be the characteristic vector of $i$th part of $\widehat{\psi}_r$, and
   let $v_i = D_r^{1/2}y_i$.
5. Return the projection onto the space spanned by the $v_i$.
```

Notice that if the expected matrix L was available, **Projection**$(D_r^{-1/2}L, 0)$ would identify the actual clusters of ψ_r, and return $P_L^{(k)}$, the projection on the column space of the expected Laplacian L, so it is not completely unnatural. The parameter τ is present to make the clustering by the **Projection** method tolerant of the error in the input.

Theorem 6 *Under the assumptions of Theorem 2, letting*

$$\tau^2 \quad = \quad c\sigma^2 k \log k \times n/64 s_{\min} \ ,$$

with probability $1-\delta-n^{-\omega(1)}$ $\widehat{M}$ *has the property that each invocation of* ***Normalized Partition***$(\widehat{M}, D_r, D_c, \tau)$ *satisfies*

$$\max_{\psi_c(u)=\psi_c(v)} |F_u - F_v| \quad \le \quad \min_{\psi_c(u)\neq\psi_c(v)} |F_u - F_v| \tag{2}$$

with constant probability strictly greater than $1/2$*.*

An MST on columns F_u satisfying (2) will connect the nodes within every part of ψ_c before connecting any two parts. As the probability of success is strictly greater than $1/2$, we can amplify the probability of success by repeating the process several times and taking the majority answer. The proof of Theorem 6 appears at the end of the Section 2.2.

2.2. Analysis

Our approach to proving Theorem 6 is to argue that the computed columns F_u are each very close to N_u, and observe that $N_u = N_v$ if and only if $\psi_c(u) = \psi_c(v)$. If $F_u \approx N_u$ for all u, then given a sufficient separation between differing N_u and N_v, the distance between F_u and F_v will reflect the distance between N_u and N_v.

Our main tool in equating F_u and N_u is the triangle inequality, bounding

$$\begin{aligned} |N_u - F_u| &\le |N_u - Q_B^{(k)}N_u| + |Q_B^{(k)}(N_u - \widehat{N}_u)| \\ |N_u - F_u| &\le |N_u - Q_A^{(k)}N_u| + |Q_A^{(k)}(N_u - \widehat{N}_u)| \end{aligned}$$

The first term is the error that the projection $Q_A^{(k)}$ or $Q_B^{(k)}$ applies to the common expected vectors N_u, which we will call the *systematic error*. The second error is the result of projecting the perturbation associated with the random rounding of each column onto the k-dimensional space, which we will call the *random error*. Lemma 8 bounds the systematic error, and Lemma 9 bounds the random error. After seeing these lemmas, we will combine them into the formal statement of Theorem 6.

Before proceeding to the heart of the analysis, we reiterate the list of conditions in Theorem 1. Entries of G are bounded by $\sigma^2 \gg (\log n)^6/n$ and when $\psi_c(u) \neq \psi_c(v)$,

$$|N_u - N_v|^2 \geq c\sigma^2 k \log k \left(\frac{n}{s_{\min}} + \frac{\log(n/\delta)}{d_{\min}} \right) \qquad (3)$$

As an immediate consequence, whenever $\psi_c(u) \neq \psi_c(v)$

$$|N_u - N_v| \geq 8\tau$$

Remark: Our analysis is conducted for the columns of N_A. The bounds apply *mutatis mutandis* for the columns of N_B.

2.2.1. Systematic Error We start our analysis of the systematic error by proving that the partition $\widehat{\psi}_r$ that **Projection** computes is a good approximation to ψ_r, in the sense that each of the parts of $\widehat{\psi}_r$ are built around a core of nodes that are only slightly perturbed.

We recall that in **Projection**, $\widehat{X} = D_r^{-1/2}\widehat{L}_A^{(k)}$, and introduce the corresponding notation $X = D_r^{-1/2} L_A$.

Lemma 7 *Under the assumptions of Theorem 2, for any fixed probability ϵ, we can choose c such that with probability at least $1-\epsilon$ each of the k nodes u_i selected in step 2a of **Projection** will satisfy*

$$|X_{u_i}^T - \widehat{X}_{u_i}^T| \quad \leq \quad \tau/2 \qquad (4)$$

Proof: The proof is essentially a counting argument, arguing that there is at most $\|L - \widehat{L}^{(k)}\|_F^2$ error to share amongst the $\widehat{X}_u^T$ and the probability of selecting a node that violates (4) is therefore small. The complete details are in the appendix. □

If each node u_i selected satisfies (4), then by marking all v within τ we mark all other rows from $\psi_r(u_i)$ that satisfy (4). Therefore, each u_i is from a different part of ψ_r, and each of the k parts of $\widehat{\psi}_r$ can be associated with a part of ψ_r. For notation's sake, we arrange $\widehat{\psi}_r$ so that $\widehat{\psi}_r(u_i) = \psi_r(u_i)$.

With this useful lemma in place, we now bound the systematic error associated with each node.

Lemma 8 *Let s_u^A be the sum of d_v of those columns in A for which $\psi_c(u) = \psi_c(v)$, and assume that $s_u^A > s_u/4$. Under the assumptions above, with probability at least $3/4$, for all u*

$$|N_u - Q_A^{(k)} N_u|^2 \quad < \quad 16\, \|L_A - \widehat{L}_A^{(k)}\|_F^2 / s_u^A$$

Proof: We start with the observation that $N_v = N_u$ when $\psi_c(v) = \psi_c(u)$, allowing us to write

$$\begin{aligned} |N_u - Q_A^{(k)} N_u|^2 &= \sum_{v \in \psi_c(u)} |N_u - Q_A^{(k)} N_u|^2 d_v / s_u^A && (5) \\ &= \sum_{v \in \psi_c(u)} |(L_A)_v - Q_A^{(k)} (L_A)_v|^2 / s_u^A && (6) \\ &\leq \|L_A - Q_A^{(k)} L_A\|_F^2 / s_u^A && (7) \end{aligned}$$

We now define a useful matrix E_A, of corresponding dimension to L_A, whose entries are

$$(E_A)_{uv} \quad = \quad d_u^{1/2} G_{\widehat{\psi}_r(u)\psi_c(v)} d_v^{1/2}$$

We can view E_A as what L_A would look like if its columns obeyed the partition ψ_c but its rows obeyed the partition $\widehat{\psi}_r$. Notice that $Q_A^{(k)} E_A = E_A$, as the columns of E_A lie in the space spanned by the v_i, the basis vectors of the projection $Q_A^{(k)}$. On the other hand, The difference $\|L_A - Q_A^{(k)} Y\|_F$ is minimized at $Y = L_A$, and so

$$\|L_A - Q_A^{(k)} L_A\|_F \quad \leq \quad \|L_A - E_A\|_F \qquad (8)$$

which can be inserted into (7),

To make the transition to $\|L_A - \widehat{L}_A^{(k)}\|_F^2$ in the numerator we consider $\|L_A - E_A\|_F^2$ as the sum of squared *row* lengths.

$$\|L_A - E_A\|_F^2 \quad = \quad \sum_u |(L_A)_u^T - (E_A)_u^T|^2$$

If $\psi_r(u) = \widehat{\psi}_r(u)$, then $(L_A)_u^T = (E_A^T)_u$, and their difference is zero. Alternately, consider two cluster centers u_i and u_j chosen in Step 2a of **Projection**, and a node u for which $\psi_r(u) = \psi_r(u_i)$, but $\widehat{\psi}_r(u) = \widehat{\psi}_r(u_j)$. For such a node, recalling that $\widehat{X} = D^{-1/2}\widehat{L}_A^{(k)}$ and $X = D^{-1/2} L_A$,

$$(L_A)_u^T = d_u^{1/2} X_{u_i}^T \quad \text{and} \quad (E_A)_u^T = d_u^{1/2} X_{u_j}^T$$

As u was associated with u_j instead of u_i, it must be that

$$|\widehat{X}_u^T - \widehat{X}_{u_j}^T| \quad \leq \quad |\widehat{X}_u^T - \widehat{X}_{u_i}^T|. \qquad (9)$$

We now start a series of inequalities with the triangle inequality (11), from which Lemma 7 yields (12). We then apply (9) to yield (13), and Lemma 7 again to yield (14).

$$\begin{aligned} & |X_{u_i}^T - X_{u_j}^T| && (10) \\ \leq\ & |X_{u_j}^T - \widehat{X}_u^T| + |\widehat{X}_u^T - X_{u_i}^T| && (11) \\ \leq\ & |\widehat{X}_{u_j}^T - \widehat{X}_u^T| + |\widehat{X}_u^T - X_{u_i}^T| + \tau/2 && (12) \\ \leq\ & |\widehat{X}_{u_i}^T - \widehat{X}_u^T| + |\widehat{X}_u^T - X_{u_i}^T| + \tau/2 && (13) \\ \leq\ & |X_{u_i}^T - \widehat{X}_u^T| + |\widehat{X}_u^T - X_{u_i}^T| + \tau && (14) \end{aligned}$$

Reorganizing and restating (14),

$$|X_{u_i}^T - X_{u_j}^T| \quad \leq \quad 2|\widehat{X}_u^T - X_{u_i}^T| + \tau \qquad (15)$$

Using our assumption that $s_u^A > s_u/4$, it is the case that

$$|X_{u_i}^T - X_{u_j}^T| \ > \ |(G_{u_j}^T - G_{u_i}^T) D^{1/2}|/4 \ > \ 2\tau$$

Inserted into (15),

$$|X_{u_i}^T - X_{u_j}^T| \quad \leq \quad 4|\widehat{X}_u^T - X_{u_i}^T| \qquad (16)$$

Recalling our definitions of $\widehat{X}$, X, and multiplying by $d_u^{1/2}$,

$$|(L_A)_u^T - (E_A)_u^T| \quad \leq \quad 4|(\widehat{L}_A^{(k)})_u^T - (L_A)_u^T| \qquad (17)$$

Finally, squaring and summing over all u,

$$\|L_A - E_A\|_F^2 \quad \leq \quad 16\|\widehat{L}_A^{(k)} - L_A\|_F^2 \qquad (18)$$

which, when substituted in (8), concludes the proof. □

2.2.2. Random Error We now move on to address the random error

$$Q_A^{(k)}(N_u - \widehat{N}_u) \quad \text{and} \quad Q_B^{(k)}(N_u - \widehat{N}_u)$$

The argument is at heart a Chernoff bound, although a little care must be taken to get it into this form.

Lemma 9 (Random Error) *Let σ^2 be an upper bound on the entries of G, and let $Q_A^{(k)}$ be a projection computed by* ***Projection****. With probability $1 - \delta$,*

$$\begin{aligned} |Q_A^{(k)}(N_u - \widehat{N}_u)|^2 \quad \leq \quad & 4k\sigma^2 \log(nk/\delta)/d_u + \\ & 8k \log^2(nk/\delta)/s_{\min}^A d_u^2 \end{aligned}$$

for all columns $u \in B$

Proof: The vectors v_i that define $Q_u^{(k)}$ are disjoint, and therefore orthogonal. As such, we can decompose $|Q_A^{(k)}(N_u - \widehat{N}_u)|^2$ into a sum of k parts, defined by the vector's projection onto each of the $v_i/|v_i|$.

$$|Q_A^{(k)}(\widehat{N}_u - N_u)|^2 = \sum_{i \leq k} (v_i^T(\widehat{N}_u - N_u))^2/|v_i|^2$$

We consider each of the k terms separately, observing that each is a sum of independent random variables with mean zero. In particular, note that

$$v_i^T(\widehat{N}_u - N_u) = y_i^T(\widehat{M}_u - M_u)/d_u$$

The entries of $\widehat{M}$ are independent $0/1$ random variables. We will apply one form of the Chernoff bound [23], which says that for a sum of $0/1$ random variables, X,

$$Pr[|X - E[X]| \geq t] \quad \leq \quad \max\{\exp(-t^2/4\mu), \exp(-t/2)\}$$

If we apply this to our sum, we see that

$$\begin{aligned} & Pr[|v_i^T(\widehat{N}_u - N_u)|^2 > t] \\ \leq \quad & Pr[y_i^T(\widehat{M}_u - M_u) > d_u t^{1/2}] \\ \leq \quad & \max\{\exp(-d_u^2 t/4\mu), \exp(-d_u t^{1/2}/2)\} \end{aligned}$$

where $\mu = E[y_i^T \widehat{M}_u]$ is bounded by

$$\begin{aligned} E[y_i^T \widehat{M}_u] \quad & \leq \quad \sum_{\widehat{\psi}_r(v)=i} d_u d_v \max_{ij} G_{ij} \\ & \leq \quad d_u \widehat{s}_i \sigma^2 \end{aligned}$$

Letting $\widehat{s}_{\min}$ to refer to the least total degree of the rows in any part of $\widehat{\psi}_r$, notice that $|v_i|^2 = \widehat{s}_i$. After instantiating

$$t = 4\sigma^2 \log(k/\delta)/d_u + 4\log^2(k/\delta)/\widehat{s}_{\min} d_u^2$$

we apply a union bound, concluding that all of the k terms are bounded by t with probability at least $1 - \delta$. In the proof of Lemma 7 we see that $\widehat{s}_u$ is at least $s_u/2$, which we use to remove the $\widehat{s}_u$ terms. □

2.2.3. Proof of Theorem 6 We now combine our bounds on the systematic and random error to prove Theorem 6.

Proof: We first recall that the probability that the random split of columns into A and B satisfies $s_u^A \geq s_u/4$ and $s_u^B \geq s_u/4$ with probability $1/2 + \varepsilon$. We conduct the rest of the proof under the assumption that this condition occurs.

Recall the triangle inequality,

$$\begin{aligned} |N_u - F_u| \quad & \leq \quad |N_u - Q_B^{(k)} N_u| + |Q_B^{(k)}(N_u - \widehat{N}_u)| \\ |N_u - F_u| \quad & \leq \quad |N_u - Q_A^{(k)} N_u| + |Q_A^{(k)}(N_u - \widehat{N}_u)| \end{aligned}$$

The systematic errors are bounded by Lemma 8 as

$$\begin{aligned} |N_u - Q_A^{(k)} N_u| \quad & \leq \quad 16\|L_A - \widehat{L}_A^{(k)}\|_F^2 / s_{\min}^A \\ |N_u - Q_B^{(k)} N_u| \quad & \leq \quad 16\|L_B - \widehat{L}_B^{(k)}\|_F^2 / s_{\min}^B \end{aligned}$$

Both $\|L_A - \widehat{L}_A^{(k)}\|_F^2$ and $\|L_B - \widehat{L}_B^{(k)}\|_F^2$ are bounded by Corollary 5 as

$$\begin{aligned} \|L_A - \widehat{L}_A^{(k)}\|_F^2 \quad & \leq \quad 256\,\sigma^2 kn \\ \|L_B - \widehat{L}_B^{(k)}\|_F^2 \quad & \leq \quad 256\,\sigma^2 kn \end{aligned}$$

with probability $1 - 4e^{-\sigma^2 n}$. Combining this with the assumption that $s_{\min}^A > s_{\min}/4$ and $s_{\min}^B > s_{\min}/4$, and taking c sufficiently large bounds

$$\begin{aligned} |N_u - Q_A^{(k)} N_u| \quad & \leq \quad c\sigma^2 kn/32 s_{\min} \\ |N_u - Q_B^{(k)} N_u| \quad & \leq \quad c\sigma^2 kn/32 s_{\min} \end{aligned}$$

with probability $1 - 4e^{-\sigma^2 n}$.

We now integrate the random error bound. This argument is a bit less direct, complicated by the involvement of randomness from both $\widehat{M}$ and the algorithm. Given a fixed $\widehat{L}_A$ satisfying $s_u^A > s_u/4$, we can view $Q_A^{(k)}$ as a random variable. The probability that any column of $\widehat{N}_B$ violates the bound of Lemma 9 for an ε fraction of the $Q_A^{(k)}$ is at most δ/ε, by the Markov inequality. Formally,

$$\begin{aligned} Pr_{\widehat{N}_B}[Pr_{Q_A^{(k)}}[violation] > \varepsilon] \quad & \leq \quad \delta/\varepsilon \\ Pr_{\widehat{N}_A}[Pr_{Q_B^{(k)}}[violation] > \varepsilon] \quad & \leq \quad \delta/\varepsilon \end{aligned}$$

For a $1 - 2\delta/\varepsilon$ fraction of the $\widehat{M}$, the bound of Lemma 9 holds for a $1 - \varepsilon$ fraction of the projections $Q_A^{(k)}, Q_B^{(k)}$. By increasing the leading constant on the bound of Lemma 9, we can decrease the value of ε to an arbitrary constant, say $1/16$.

Combining the two bounds, notice that if we insert the assumption $\sigma^2 \gg (\log n)^6/n$, the systematic error exceeds the second term in the bound of Lemma 9. Each execution of **Normalized Partition** therefore satisfies, for a large c

$$|N_u - F_u|^2 \quad \leq \quad c\sigma^2(nk/s_{\min} + k\log(n/\delta)/d_u)/16 \quad (19)$$

for all u, when

1. The split of degrees is balanced, (probability $3/4$)
2. Lemma 7 holds for $\widehat{N}_A$ and $\widehat{N}_B$, (prob. $(15/16)^2$)
3. Lemma 9 holds for $Q_A^{(k)}$ and $Q_B^{(k)}$. (prob. $14/16$)

The probability all occur is

$$3/4 - 3/4(1-(15/16)^2) - 3/4(1-14/16) \quad > \quad 1/2\ .$$

As when $\psi(u)_c \neq \psi_c(v)$ the separation $|N_u - N_v|^2$ is at least sixteen times the right hand side of (19),

$$|N_u - F_u| \quad \leq \quad \min_{\psi_c(u)\neq\psi_c(v)} |N_u - N_v|/4\ . \quad (20)$$

Using the triangle inequality,

$$\begin{aligned} \max_{\psi_c(u)=\psi_c(v)} |F_u - F_v| \quad &\leq \quad |N_u - F_u| + |N_v + F_v| \\ &\leq \quad \min_{\psi_c(u)\neq\psi_c(v)} |N_u - N_v|/2 \\ &\leq \quad \min_{\psi_c(u)\neq\psi_c(v)} |F_u - F_v| \end{aligned}$$

and we conclude the proof. □

3. Extensions

This paper proposes a natural transformation to deal with skewed degree distributions in spectral analysis. In order to provide rigorous proofs, we resorted to a synthetic projection inspired by [21]. An interesting question is to explore whether we can prove that the eigenvector projection performs well. Also, the assumptions in this paper imply a lower bound of $\log^6 n$ on the minimum degree of the graph. Analyzing sparser graphs seems to require stronger matrix bounds than are available now.

Another important direction to explore involves the deconditioning of the algorithm. The partitioning of the graph in Step 1 of the algorithm was prompted by the need to decondition the computed projection from the random vectors themselves. Unfortunately, the partitioning itself is sensitive to the "smoothness" of the degree distribution. When there are very few nodes of very high degree, with some probability all of them will be placed in only one of the partitions, and hence the partitions are likely to be unbalanced. A different approach to avoid conditioning is to remove each node from the vertex set before analyzing it. In effect, for computing the projection for column u, we remove u from the vertex set and then run **Projection** on the remaining data. Though this approach has the potential of being robust against the aforementioned degree distributions, proving the analogue of Lemma 7 for this method seems hard.

Finally, it would be interesting to apply this normalization technique and analysis to other, less synthetic domains. For example, collaborative filtering experiences a tremendous skew in the popularity of items that users rate. At the same time, the utility of such systems is to discover less common items of interest, as common interesting items are usually easily discovered. This particular application seems like it could benefit substantially from normalization, both in theory and in practice.

References

[1] D. Achlioptas, A. Fiat, A. Karlin, F. McSherry. Web search via hub synthesis. *FOCS* 2001.

[2] D. Achlioptas and F. McSherry, Fast Computation of Low Rank Matrix Approximations, *STOC*, 2001.

[3] N. Alon, M. Krivelevich, B. Sudakov, Finding a large hidden clique in a random graph, *SODA* 1998.

[4] N. Alon, M. Krivelevich, V. Vu, On the concentration of eigenvalues of random symmetric matrices, *Tech. Report 60, Microsoft Research*, 2000.

[5] C. J. Alpert and S. Z. Yao, Spectral Partitioning: The More Eigenvectors, the Better, *UCLA CS Dept. Technical Report*, #940036, October 1994.

[6] Y. Azar, A. Fiat, A. Karlin, F. McSherry, J. Saia Spectral Analysis of Data *STOC*, 2001.

[7] R. B. Boppana. Eigenvalues and graph bisection: an average-case analysis. *FOCS*, 1987.

[8] F. R. K. Chung, L. Lu, V. Vu. The spectra of random graphs with given expected degrees *PNAS* 100, no. 11, (2003), 6313–6318.

[9] A. Condon and R. M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures and Algorithms*, 18:2, 116–140, 2001.

[10] C. Davis and W. Kahan, The Rotation of Eigenvectors by a Perturbation 3. SIAM Journal on Numerical Analysis, Vol 7, 1970, 1–46.

[11] P. Husbands, H. Simon, and C. Ding, On the Use of the Singular Value Decomposition for Text Retrieval. Proc. of SIAM Comp. Info. Retrieval Workshop.

[12] I. Dhillon, Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning. *SIGKDD*, 2001.

[13] H. Zha, X. He, C. Ding, M. Gu and H. Simon, Bipartite Graph Partitioning and Data Clustering. *CIKM*, 2001.

[14] P. Drineas, R. Kannan, A. Frieze, S. Vempala, and V. Vinay, Clustering in large graphs and matrices. SODA, 1999.

[15] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, Indexing by latent semantic analysis. Journal of the Society for Information Science, Vol 41, 1990, 391–407.

[16] Z. Furedi and J. Komlos, The eigenvalues of random symmetric matrices, *Combinatorica* 1(3), 233–241 (1981).

[17] G. Golub, C. Van Loan (1996), Matrix computations, third edition, *The Johns Hopkins University Press Ltd., London*

[18] K. S. Jones, S. Walker, S. E. Robertson Probabilistic Model of Information Retrieval: Development and Status, *Information Processing and Management*, 1998.

[19] R. Kannan, S. Vempala, A. Vetta On clusterings-good, bad and spectral. *FOCS*, 2000.

[20] J. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM, Vol 46, 1999, 604–632.

[21] F. McSherry. Spectral Partitioning of Random Graphs. *FOCS*, 2001.

[22] M. Mihail, C.Papadimitriou. On the Eigenvalue Power Law. *RANDOM* 2002.

[23] R. Motwani and P. Raghavan. Randomized Algorithms. Cambridge University Press, 1995.

[24] A. Ng and M. Jordan and Y. Weiss. On spectral clustering: Analysis and an algorithm, *NIPS* 2001.

[25] L. Page and S. Brin, PageRank, an Eigenvector based Ranking Approach for Hypertext. *SIGIR*, 1998

[26] C. H. Papadimitriou, H. Tamaki, P. Raghavan and S. Vempala, Latent Semantic Indexing: A Probabilistic Analysis. *PODS*, 1998.

[27] R. Lempel, S. Moran. SALSA: the stochastic approach for link-structure analysis. *TOIS* 19(2): 131-160 (2001)

[28] R. Lempel, S. Moran. Rank Stability and Rank Similarity of Link-Based Web Ranking Algorithms in Authority Connected Graphs. *Information Retrieval*, special issue on Advances in Mathematics/Formal Methods in Information Retrieval.

[29] M. E. J. Newman, S. H. Strogatz, and D. J. Watts., Random graphs with arbitrary degree distribution and their application. *Physical Review* E 64, 026118 (2001).

[30] D. A. Spielman, S. H. Teng, Spectral Partitioning Works: Planar graphs and finite element meshes. *FOCS*, 1996.

[31] G. W. Stewart, J. G. Sun, Matrix perturbation theory. Academic Press Inc., Boston MA, 1990

[32] E. Wigner, On the distribution of the roots of certain symmetric matrices, *Annales of Math.* Vol. 67, 325–327, 1958.

[33] J. Zobel, A. Moffat. Exploring the Similarity Space, *SIGIR*. 32(1), 18–34, 1998.

Appendix : Proof of Lemma 7

Proof: Recall that we wanted to prove that each node selected as cluster center satisfies

$$|X_{u_i}^T - \widehat{X}_{u_i}^T| \leq \tau/2$$

with probability at least $15/16$.

Let us call any node that satisfies this bound "good"; other rows will be called "bad". Notice that if at each step we do choose a good node, then by using radius τ we will mark all the good rows from the same part as u, and no good rows from any other part (as the centers are assumed to be separated by 2τ). If we were to chose good rows only, the proof would be complete.

So, let us look at the probability of choosing a bad node at a particular step. Recall that we choose rows proportional to their degree. By our definition, the total degree of bad rows is bounded as

$$\begin{aligned} \sum_{u\in BAD} d_u &\leq \sum_{u\in BAD} 4d_u|X_u^T - \widehat{X}_u^T|^2/\tau^2 \\ &= \sum_{u\in BAD} 4|(L_A)_u^T - (\widehat{L}_A)_u^T)|^2/\tau^2 \\ &\leq 4\|L_A - \widehat{L}_A^{(k)}\|_F^2/\tau^2 \end{aligned}$$

Combining our assumption on the size of τ with the bound of Corollary 5, we see that

$$\tau \geq c/2^{14} \times \|L_A - \widehat{L}_A^{(k)}\|_F^2 \log k/s_{\min}$$

We can choose c such that

$$\sum_{u\in BAD} d_u \leq s_{\min}/16\log k$$

However, notice that in the first step we have at least $ks_{\min}$ total row degree, and so the probability of choosing a bad row is at most $(16k\log k)^{-1}$. Indeed, at any step i at which we have not chosen a bad row, there is still $(k-i+1)s_{\min}$ total degree, as we have only marked rows from $i-1$ parts. The probability of selecting a bad row at the ith step is therefore at most

$$Pr[\text{select bad } u_i] \leq (16(k-i+1)\log k)^{-1}$$

If we now take a union bound, we see that

$$\begin{aligned} Pr[\text{any bad}] &\leq \frac{\sum_{i\leq k}(k-i+1)^{-1}}{16\log k} \\ &\approx \epsilon \end{aligned}$$

As we choose only good rows, and at each step mark all good rows associated with a particular part, each selection must be from a different part. □