Gossip PCA

Satish Babu Korada
Goldman Sachs
New York, NY 10004
satishbabu.k@gmail.com

Andrea Montanari
Electrical Engineering and Statistics Departments
Stanford University
Stanford, CA 94305
montanari@stanford.edu

Sewoong Oh
EECS Department
Massachusetts Institute of Technology
Cambridge, MA 02139
swoh@mit.edu

ABSTRACT

Eigenvectors of data matrices play an important role in many computational problems, ranging from signal processing to machine learning and control. For instance, algorithms that compute positions of the nodes of a wireless network on the basis of pairwise distance measurements require a few leading eigenvectors of the distances matrix. While eigenvector calculation is a standard topic in numerical linear algebra, it becomes challenging under severe communication or computation constraints, or in absence of central scheduling. In this paper we investigate the possibility of computing the leading eigenvectors of a large data matrix through gossip algorithms.

The proposed algorithm amounts to iteratively multiplying a vector by independent random sparsification of the original matrix and averaging the resulting normalized vectors. This can be viewed as a generalization of gossip algorithms for consensus, but the resulting dynamics is significantly more intricate. Our analysis is based on controlling the convergence to stationarity of the associated Kesten-Furstenberg Markov chain.

Categories and Subject Descriptors

C.2.4 [**Computer-Communication Networks**]: Distributed Systems—*Distributed applications*

General Terms

Algorithms, Performance

1. INTRODUCTION AND OVERVIEW

Consider a system formed by n nodes with limited computation and communication capabilities, and connected via the complete graph K_n. To each edge (i, j) of the graph is associated the entry M_{ij} of an $n \times n$ symmetric matrix M. Node i has access to the entries of M_{ij} for $j \in \{1, \dots, n\}$. An algorithm is required to compute the eigenvector of M corresponding to the eigenvalue with the largest magnitude. Denoting by $u \in \mathbb{R}^n$ the eigenvector, each node i has to compute the corresponding entry u_i. The eigenvector u is often called the *principal component* of M, and analysis methods that approximate a data matrix by its leading eigenvectors are referred to as principal component analysis [19].

Eigenvector calculation is a key step in many computational tasks, e.g. dimensionality reduction [29], classification [17], latent semantic indexing [7], link analysis (as in PageRank) [6]. The primitive developed in this paper can therefore be useful whenever such tasks have to be performed under stringent communication and computation constraints. As a stylized application, consider the case in which the nodes are n wireless hand-held devices (for related commercial products, see [1, 2, 3]). Accurate positioning of the nodes in indoor environments is difficult through standard methods such as GPS [26]. Because of intrinsic limitation of GPS and of roof scattering, indoor position uncertainty can be of 10 meters or larger, which is too much for locating a room in a building. An alternative approach consists in measuring pairwise distances through delay measurements between the nodes and reconstructing the nodes positions from such measurements (obviously this is possible only up to a global rotation or translation). Positions indeed can be extracted from the matrix of square distances by computing its three leading eigenvectors (after appropriate centering) [24]. This method is known as multidimensional scaling, and we will use it as a running example throughout this paper.

A simple centralized method for computing the eigenvector is to collect all the matrix entries at one special node, say node i, to perform the eigenvector calculation there and then flood back its entries to each node. This centralized approach has several disadvantages. It requires communicating n^2 real numbers through the network at the beginning of execution, and puts a large memory, computation and communication burden on node i. It is also very fragile to failure or Byzantine behavior of i.

The next simplest idea is to use some version of the *power method*. A decentralized power method would proceed by synchronized iterations through the network. At t-th iteration, each node keeps a running estimate $x_i^{(t)}$ of the leading eigenvector. This is updated by letting $x_i^{(t+1)} = \sum_{j=1}^{n} M_{ij} x_j^{(t)}$. If M has strong spectral features (in particular, if the two largest eigenvalues are not close) these estimates will converge rapidly. On the other hand, each iteration requires $(n-1)$ real numbers to be transmitted to each node, and

n sums and multiplications to be performed at the node. In other words, the node capabilities have to scale with the network size. This problem becomes even more severe for wireless devices, which are intrinsically interference-limited. Within the power method approach, n^2 communications have to be scheduled at each time thus requiring significant bandwidth. Finally, the algorithm requires complete synchronization of the n^2 communications and is fragile to link failures (which can be quite frequent e.g. due to fading).

A simple and yet powerful idea that overcomes some of these problems is sparsification. Throughout the paper, we say that $S \in \mathbb{R}^{n\times n}$ is a *sparsification* of M if it is obtained by setting to 0 some of the entries of M and (eventually) rescaling the non-zero entries. A sparsification is useful if most of its entries are zero, and yet the resulting matrix has a leading eigenvector close to the original one. Given a sparsified matrix S, power method can be applied by $x_i^{(t+1)} = \sum_{j=1}^n S_{ij}x_j^{(t)}$. If S has d-nonzero entries per row, each node needs to communicate d real numbers, and to perform d sums and multiplications. For wireless devices, the badwidth scales at most like nd.

In [4] Achlioptas and McSherry showed that a sparsification can be constructed such that

$$\|M - S\|_2 \le \theta \, \|M\|_2 \, , \tag{1}$$

with only $d = O(1/\theta^2)$ non-zero entries per row. The inequality (1) immediately implies that computing the leading eigenvector of S, yields an estimator $\widehat{u}$ that satisfies $\|\widehat{u} - u\| \le 2\theta$. (Here and below, for $v, w \in \mathbb{R}^m$, v^* denotes its transpose and $\langle v, w\rangle = v^*w$ denotes the scalar product of two vectors. Let $\|v\| = \langle v, v\rangle$ denote its Euclidean –or ℓ_2– norm, i.e. $\|v\|^2 \equiv \sum_{i=1}^n v_i^2$. For a matrix A, $\|A\|_2$ denotes its ℓ_2 operator norm, i.e. $\|A\|_2 \equiv \sup_{v\neq 0} \|Av\|/\|v\|$.) The construction of [4] is based on random sampling. Each entry of M is set to 0 independently with a given probability $1 - p = 1 - d/n$. Non-zero entries are then rescaled by a factor $1/p$. The bound (1) is proved to hold with high probability with respect to the randomness in the sparsification.

While this approach is simple and effective, it still presents important shortcomings: (*i*) For a fixed per node complexity which scales like $1/\theta^2$, this procedure achieves precision θ: can one achieve a better scaling? (*ii*) A fixed subnetwork G of the complete graph (corresponding to the sparsity pattern of S) needs to be maintained through the whole process. This can be challenging in the presence of fading or of node failures/departures. (*iii*) The target precision is to be decided at the beginning of the process, when the sparsification is constructed.

In this paper we use sparsification as a primitive and propose a new way to exploit its advantages. Roughly speaking at each round t a new independent sparsification $S^{(t)}$ of M is produced. Estimates of the leading eigenvector are generated by applying $S^{(t)}$, i.e. through

$$x^{(t)} = S^{(t)}x^{(t-1)} \, , \tag{2}$$

and then averaging across iterations $\widehat{u}^{(t)} \propto \sum_{\ell\le t} x^{(\ell)}/\|x^{(\ell)}\|$. We will refer to this algorithm as Gossip PCA. In the limit case in which $S^{(t)}$ are in fact deterministic and coincide with a fixed S, the present scheme reduces to the previous one. However, general independent random sparsifications $S^{(t)}$ can model the effect of fading, short term link failures, node departures. (While complete independence is a simplistic model for these effects, it should be possible to include short time-scale correlations in our treatment.) Finally, the use of truly random, independent sparsifications might be a choice of the algorithm designer.

Does the time-variability of $S^{(t)}$ deteriorate the algorithm precision? Surprisingly, the opposite turns out to be true: Using independent sparsifications appears to benefit accuracy by effectively averaging over a larger sample of the entries of M. As an example consider the sparsification scheme mentioned above, namely each entry of $S^{(t)}$ is set to 0 independently with a fixed probability $1 - p$. Then, with respect to the total per-node computation and communication budget, scaling of the ℓ_2 error $\|\widehat{u} - u\|$ remains roughly the same as in the time-independent case (see Section 3). Remarkably, the way optimal accuracy is achieved is significantly different from the one that is optimal within the time-independent case. In the latter case it is optimal to invest resources in the densest possible sparsification S, and then iterate it a few times. Within the present approach, one should rather use much sparser matrices $S^{(t)}$ and iterate the basic update (2) many more times. The use of sparser subnetworks is advantageous both for robustness and the overhead of maintaining/synchronizing such networks.

Our main analytical result is an error bound for the time-dependent iteration (2), that takes the form

$$\|\widehat{u}^{(t)} - u\| \le C\left(\theta/\sqrt{t} \, + \, \theta^2 \log(1/\theta)^2\right) \, , \tag{3}$$

with a constant C explicitly given below. Notice that, for t large enough, this yields an error roughly of size θ^2. While using the same number of communications per node, this is significantly smaller than the error θ obtained by computing the leading eigenvector of a single sparsification.

The upper bound (3) holds under the following three assumptions: (*i*) $\|M - S^{(\ell)}\|_2 \le \theta\|M\|_2$ for all $\ell \le t$; (*ii*) $\mathbb{E}(S^{(\ell)}) = M$; (*iii*) $S^{(\ell)}$ invertible for all ℓ. Further it is required that the initial condition satisfies $\|x^{(0)} - u\| \le C\theta$. This can be generated by iterating a fixed sparsification (say $S^{(1)}$) for a modest number of iterations (roughly $\log(1/\theta)$). Numerical simulations and heuristic arguments further suggest that the last assumption is actually a proof artifact and not needed in practice (see further discussion in Section 2.2).

The rest of the paper is organized as follows: Section 2 provides a formal description of our algorithms and of our general performance guarantees. In Section 3 we discuss implications of our analysis in specific settings. Section 4 reviews related work on randomized low complexity methods. Section 5 describes the proof of our main theorem. This leverages on the theory of products of random matrices, a line of research initiated by Furstenberg and Kesten in the sixties [15], with remarkable applications in dynamical systems theory [25]. The classical theory focuses however on matrices of fixed dimension, in the limit of an infinite number of iterations, while here we are interested in high-dimensional (large n) applications. We need therefore to characterize the tradeoff between dimensions and number of iterations.

In Section 6, we provide the proof of the technical lemmas used in the main proof. Finally, Section 7 discusses extending our algorithm to estimate the largest eigenvalues and provides a general performance guarantee.

2. MAIN RESULTS

In this section, we spell out the algorithm execution and state the main performance guarantee.

2.1 Algorithm

As mentioned in the previous section $M \in \mathbb{R}^{n\times n}$ is a symmetric matrix, with eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$. Without loss of generality, we assume that the largest eigenvalue λ_1 is positive. Further, we assume $\lambda_1 > |\lambda_2|$ strictly. We will also write $\lambda \equiv \lambda_1$ and u for the corresponding eigenvector. We assume to have at our disposal a primitive that outputs a random sparsification S of M. A sequence of independent such sparsifications will be denoted by $\{S^{(1)}, S^{(2)}, \dots\}$. In the next two paragraphs we describe a centralized version of the algorithm, and then the fully decentralized one.

2.1.1 Centralized algorithm

The system is initialized to a vector $x^{(0)} \in \mathbb{R}^n$. Then we iteratively multiply the i.i.d. sparsifications $S^{(1)}, S^{(2)}, \dots$ to get a sequence of vectors $x^{(1)}, x^{(2)}, \dots$. After t iterations, our estimate for the leading eigenvector u is

$$\widehat{u}^{(t)} = c(t) \sum_{s=1}^{t} \frac{x^{(s)}}{\|x^{(s)}\|}\,, \tag{4}$$

with $c(t)$ the appropriate normalization to ensure $\|\widehat{u}^{(t)}\| = 1$.

Note that, even after normalization, there is a residual sign ambiguity: both u and $-u$ are eigenvector. When in the following we write that $\widehat{u}^{(t)}$ approximate u within a certain accuracy, it is understood that $\widehat{u}^{(t)}$ does in fact approximate the closest of u and $-u$. A more formal resolution of this ambiguity uses the projective manifold define in Section 5.

2.1.2 Decentralized algorithm

The algorithm described so far uses the following operations:
(*i*) *Multiplying vector* $x^{(t-1)}$ *by* $S^{(t)}$, cf. Eq. (2). If $S^{(t)}$ has dn non-zero elements, this requires $O(d)$ operations per node per round.
(*ii*) *Computing the normalizations* $\|x^{(1)}\|$, $\|x^{(2)}\|$,…,$\|x^{(t)}\|$. Since $\|x^{(\ell)}\|^2 = \sum_{i=1}^n (x_i^{(\ell)})^2$, this task can be performed via a standard gossip algorithm. This entails an overhead of $\log(1/\varepsilon)$ per node per iteration for a target precision ε. We will neglect this contribution in what follows.
(*iii*) *Averaging normalized vectors across iterations,* cf. Eq. (4). Since node i keeps the sequence of estimates $x_i^{(1)}$,…, $x_i^{(t)}$, this can be done without communication overhead, with $O(1)$ computation per node per iteration.

Finally the normalization constant $c(t)$ in Eq. (4) needs to be computed. This amounts to computing the norm of the vector on the right hand side of Eq. (4), which is the same operation as in step (2) (but has to be carried out only once). From this description, it is clear that operation (1) (matrix-vector multiplication) dominates the complexity and we will focus on this in our discussion below and in Section 3.

2.2 Analysis

The algorithm design/optimization amounts to the choice of number of iterations t and the the sparsification method, which produces the i.i.d. matrices $\{S^{(\ell)}\}$. The latter is characterized by two parameters: θ which bounds the sparsification accuracy as per Eq. (1), and d, the average number of non-zero entries per row, which determines its complexity.

The trade-off between d and θ depends on the sparsification method and will be further discussed in the next section. Our main result bounds the error of the algorithm in terms of θ, t and of a characteristic of the matrix M, namely the ratio of the two largest eigenvalues $l_2 = |\lambda_2|/\lambda$. The proof of this theorem is presented in Section 5.

THEOREM 2.1. *Let $\{S^{(\ell)}\}_{\ell\geq 1}$ be a sequence of i.i.d. $n\times n$ random matrices such that $\mathbb{E}[S^{(\ell)}] = M$, $\|S^{(\ell)} - M\|_2 \leq \theta\|M\|_2$, $S^{(\ell)}$ is almost surely non-singular, and there is no proper subspace $V \subseteq \mathbb{R}^n$ such that $S^{(\ell)}V \subseteq V$ almost surely. Further, let $x^{(0)} \in \mathbb{R}^n$ be such that $\|x^{(0)} - u\| \leq \theta/(1-l_2)$ for the leading eigenvector of u. Let the eigenvector estimates be defined as per Eq. (2) and (4). Finally assume $\theta \leq (1/40)(1-l_2)^{3/2}$ and let $l_2 \equiv |\lambda_2|/\lambda$.*

Then, with probability larger than $1 - \max(\delta, 16/n^2)$,

$$\|\widehat{u}^{(t)} - u\| \leq \frac{18\theta}{(1-l_2)\sqrt{t\delta}} + 12\Big(\frac{\theta\log(1/\theta)}{(1-l_2)}\Big)^2\,. \tag{5}$$

The assumption on the samples $\{S^{(\ell)}\}_{\ell\geq 0}$ are rather mild. The matrix whose eigenvector we are computing is the expectation of $S^{(\ell)}$, the variability of $S^{(\ell)}$ is bounded in operator norm, and finally the $S^{(\ell)}$ are sufficiently random (in particular the do not share an eigenvector *exactly*). The latter can be ensured by adding arbitrarily small random perturbation to $S^{(\ell)}$.

At first sight, the assumption $\|x^{(0)} - u\| \leq \theta/(1-l_2)$ on the initial condition might appear unrealistic: the algorithm requires as input an approximation of the eigenvector u. A few remarks are in order. *First,* the accuracy of the output, see Eq. (5), is dramatically higher than on the input for $t = \Omega(1/\theta^2)$. In the following section, we will see that this is indeed the correct scaling of t that achieves optimal perormance. *Second,* numerical simulations show clearly that, for $\mathbf{x}^{(t)} = x^{(t)}/\|x^{(t)}\|$, the condition $\|\mathbf{x}^{(t)} - u\| \leq \theta/(1-l_2)$ is indeed satisfied after a few iterations. The heuristic argument is that the leading eigenvectors of $S^{(1)}$, $S^{(2)}$, $\dots S^{(t)}$ are roughly aligned with u, and their second eigenvalues are significantly smaller. Hence the scalar product $Z_t \equiv \langle u, \mathbf{x}^{(t)}\rangle$ behaves approximately as a random walk with drift pushing out of $Z_t = 0$. Even if $Z_0 = 0$, random fluctuations produce a non-vanishing Z_t, and the drift amplify this fluctuation exponentially fast. The arguments in Section 5 further confirm this heuristic argument. For instance we will prove that the set $\|\mathbf{x}^{(t)} - u\| \leq \theta/(1-l_2)$ is absorbing, in the sense that starting from such a set, the power iteration keeps $\mathbf{x}^{(t)}$ in the same set. On the other hand, starting from any other point, there is positive probability of reaching the absorbing set. Finally, further evidence is provided by the fact that random initialization is sufficient for the eigenvalue estimation as proved in Section 7.

As an example, we randomly generated a marix M and computed $\mathbf{x}^{(t)} = x^{(t)}/\|x^{(t)}\|$ according to (2) using random sparsifications with dn entries. Let $\tau = \arg\min_t\{\|\mathbf{x}^{(t)}_{\rm rand} - \mathbf{x}^{(t)}_{\rm u}\| \leq 0.001\}$. The subscript denotes two different initializations: $x^{(0)}_{\rm rand}$ is initialized with i.i.d Gaussian entries, and $x^{(0)}_u = u$. The following result illustrates that after a few iterations $t = O(\log(1/\theta))$, $\mathbf{x}^{(t)}$ achieves error of order θ with $d = O(1/\theta^2)$ operations per node per round.

d	40	80	160	320
τ	5.1	4.8	4.2	3.7
$\|\mathbf{x}^{(\tau)}_{\rm rand} - u\|$	0.1110	0.0761	0.0521	0.0329

Finally, constructing a rough approximation of the leading eigenvector is in fact an easy task by multiplying the same sparsification $S^{(0)}$ a few times. This claim is made precise by the following elementary remark.

REMARK 2.2. *Assume that $x^{(0)}$ have i.i.d. components $\mathsf{N}(0, 1/n)$, and define $x^{(t)} = S^{(t)}x^{(t-1)}$ where for $t \leq t_*$, $S^{(t)} = S$ is time independent and satisfies $\|S - M\|_2 \leq (\theta^2/2(1-l_2))\|M\|_2$. If $t_* \geq 3\log(n/\theta)/(1-l_2-\theta)$, then $\|\mathbf{x}^{(t_*)} - u\| \leq \theta/(1-l_2)$ with probability at least $1 - 1/n^2$.*

The content of this remark is fairly intuitive: the principal eigenvector of S is close to u, and the component of $x^{(t)}$ along it grows exponentially faster than the other components. A logarithmic number of iterations is then sufficient to achieve the desired distance from u.

Finally, consider the assumption $\mathbb{E}[S^{(\ell)}] = M$. In practice, it might be difficult to produce unbiased sparsifications: does Theorem 2.1 provide any guarantee in this case? The answer is clearly affirmative. Let $\mathbb{E}[S^{(\ell)}] = M'$ and assume $\|M - M'\|_2 \leq \theta'\|M\|_2$. Then, it follows immediately from (5) that

$$\|\widehat{u}^{(t)} - u\| \leq \frac{18\theta}{(1-l_2)\sqrt{t\delta}} + 12\Big(\frac{\theta\log(1/\theta)}{(1-l_2)}\Big)^2 + \frac{2\theta'}{1-l_2}\,,$$

In other words the eigenvector approximation degrades gracefully with the quality of the sparsification.

3. EXAMPLES AND APPLICATIONS

In this section we apply our main theorem to specific settings and point out possible extensions.

3.1 Computation-accuracy tradeoff

As mentioned above, Theorem 2.1 characterizes the scaling of accuracy with the quality of the sparsification procedure. For the sake of simplicity, we will consider the case in each entry of M is set to 0 independently with a fixed probability $1 - d/n$, and non-zero entries are rescaled. In other words $S_{ij} = (n/d)M_{ij}$ with probability (d/n), and $S_{ij} = 0$ otherwise. This scheme was first analyzed in [4], but the estimate only holds for $d \geq (8\log n)^4$. This condition was refined in [22]. Noting that for $d > \log n$ the maximum number of entries per row is of order d, the latter gives

$$\|M - S\|_2 \leq (C/\sqrt{d})\|M\|_2 \equiv \theta\|M\|_2\,.$$

In other words i.i.d. sparsification of the entries yields $\theta = O(1/\sqrt{d})$. Further, denoting the total complexity per node by χ, we have $\chi \sim td$ either in terms of communication or of computation.

In order to compute a computation-accuracy tradeoff we need to link the accuracy to t and θ. Let us first consider the case in which a single sparsification S is used by letting $x^{(t)} = Sx^{(t-1)}$ and $\widehat{u}^{(t)} = x^{(t)}/\|x^{(t)}\|$. This procedure converges exponentially fast to the leading eigenvector of S which in turn satisfies $\|\widehat{u}^{(\infty)} - u\| \leq 2\theta \leq C'(1/\sqrt{d})$. Therefore if we denote by $\Delta_{\rm PM} \equiv \|\widehat{u}^{(t)} - u\|$ the corresponding error after t iterations, we have

$$\Delta_{\rm PM} \sim \theta + e^{-at}\,,$$

where we deliberately omit constants since we are only interested in capturing the scaling behavior.

Now we assume that we have a limit on the total complexity $\chi \sim td$, and minimize the error $\Delta_{\rm PM}$ under this resources constraint, using the relation $\theta \sim 1/\sqrt{d}$. A simple calculation shows that the smallest error is achieved when $t = \Theta(\log\chi)$ yielding

$$\Delta_{\rm PM} \sim \sqrt{(\log\chi)/\chi}\,. \tag{6}$$

Next consider the algorithm developed in the present paper, GOSSIP PCA. The only element to be changed in our analysis is the relation between accuracy and the parameters θ and t. From Theorem 2.1 we know that our estimator achieves error $\Delta_{\rm Gossip} = \|\widehat{u}^{(t)} - u\|$ that scales as

$$\Delta_{\rm Gossip} \sim \theta/\sqrt{t} + \big(\theta\log(1/\theta)\big)^2\,,$$

where again we omit constants. It is straightforward to minimize this expression under the constraints $\chi \sim td$, and $\theta \sim 1/\sqrt{d}$. The best scaling is achieved when $t = \Theta(\sqrt{\chi}/(\log\chi)^2)$ and $\theta = \Theta(1/(\chi^{1/4}\log\chi))$ yielding

$$\Delta_{\rm Gossip} \sim 1/\sqrt{\chi}\,. \tag{7}$$

Comparing (6) and (7), the scaling of the error with the per-node computation and communication remains roughly the same up to a logarithmic factor. Surprisingly, the way the best accuracy is achieved is significantly different. In the time-independent case (the standard power method), it is optimal to invest a lot of resources in one iteration with a dense matrix S that has $d = \Theta(\chi/\log\chi)$ non-zero entries per row. In return, only a few iterations $t = \Theta(\log\chi)$ are required. Within the proposed time-dependent gossip approach, one should rather use a much sparser matrices $S^{(\ell)}$ with $d = \Theta(\sqrt{\chi}(\log\chi)^2)$ non-zero entries per row and use a larger number of iterations $t = \Theta(\sqrt{\chi}/(\log\chi)^2)$.

To illustrate how the two gossip algorithms compare in practice, we present results of a numerical experiment from the positioning application. From 1000 nodes placed in the 2-dimensional unit square uniformly at random, we define the matrix of squared distances. Let p_i be the position of node i, then $D_{ij} = \|p_i - p_j\|^2$. After a simple centering operation, the top two eigenvectors reveal the position of the nodes up to a rigid motion (translation and/or rotation) [24]. We can extend the gossip algorithms to estimate the first two eigenvectors as explained in Section 3.3. Let the columns of $U \in \mathbb{R}^{1000\times 2}$ be the first two eigenvectors and $\|\cdot\|_F$ be the Frobenius norm of a matrix such that $\|A\|_F^2 = \sum_{i,j} A_{ij}^2$. Denote by $\Delta(d) = (1/\sqrt{2})\|U - \widehat{U}\|_F$ the resulting error for a particular choice of d.

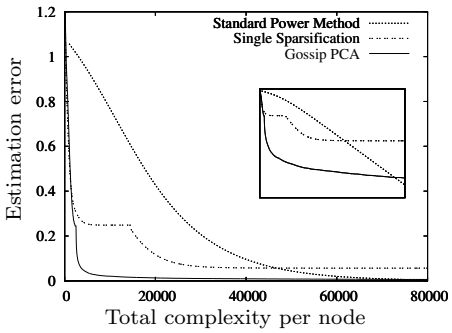

Figure 1: Eigenvector estimation error against complexity. In the inset the result is plotted in log-scale.

To simulate a simple gossip setting with constrained communication, we allow d to be either 50 or 500. For the two gossip algorithms and for each value of the total complexity χ, we plot the minimum error achieved using one of the two allowed communication schemes: $\min_{d\in\{50,500\}} \Delta(d)$. For comparison, performance of the power method on complete dense matrices is also shown (see Section 1). As expected from the analysis, GOSSIP PCA achieved smaller error with sparse matrices ($d = 50$) for all values of χ. When a single sparsification is used, there is a threshhold at $\chi = 14500$, above which a dense matrix ($d = 500$) achieved smaller error. Notice a discontinuity of the derivative at the threshold.

3.2 Comparison with gossip averaging

Gossip methods have been quite successful in computing symmetric functions of data $\{x_i^{(0)}\}_{1\le i\le n}$ available at the nodes. The basic primitive in this setting is a procedure computing the average $\sum_{i=1}^n x_i^{(0)}/n$. This algorithm shares similarities with the present one. One recursively applies independent random matrices $P^{(1)}$, $P^{(2)}$, ... according to:

$$x^{(t)} = P^{(t)} x^{(t-1)}\,, \tag{8}$$

where $P^{(t)}$ is the matrix that averages entries $i(t)$ and $j(t)$ of $x^{(t)}$ (in other words it is the identity outside a 2×2 block corresponding to coordinates $i(t)$ and $j(t)$).

It is instructive to compare the two problems. In the case of simple averaging, one is interested in approximating the action of a projector P, namely the matrix with all entries equal to $1/n$. In eigenvector calculations the situation is not as simple, because the matrix of interest M is not a simple projector. In both cases we approximate this action by products of i.i.d. random matrices whose expectation matches the matrix of interest. However in averaging, the leading eigenvector of P is known *a priori*, it is the constant vector $u = (1/\sqrt{n}, \dots, 1/\sqrt{n})$. As a consequence, sparsifications $P^{(t)}$ can be constructed in such a way that $P^{(t)}u = u$ with probability 1.

Reflecting these differences, the behavior of the present algorithm is qualitatively different from gossip averaging. Within the latter $x^{(t)}$ converges asymptotically to the constant vector, whose entries are equal to $\sum_{i=1}^n x_i^{(0)}/n$. The convergence rate depends on the distribution of the sparsification $P^{(t)}$. In GOSSIP PCA, the sequence of normalized vectors $x^{(t)}/\|x^{(t)}\|$ does not converges to a fixed point. The distribution of $x^{(t)}/\|x^{(t)}\|$ instead converges to a non-trivial stationary distribution whose mean is approximated by $\widehat{u}^{(t)}$. An important step in the proof of Theorem 2.1 consists in showing that the mean of this distribution is much closer to the eigenvector than a typical vector drawn from it.

3.3 Extensions

It is worth pointing out some extensions of our results, and interesting research directions:

More than one eigenvector. In many applications of interest, we need to compute r leading eigenvectors, where r is larger than one, but typically a small number. In the case of positioning wireless devices, r is consistent with the ambient dimensions, hence $r = 3$. As for the standard power iteration, the algorithm proposed here can be generalized to this problem. At iteration t, the algorithm keeps track of r orthonormal vectors $x^{(t)}(1), \dots x^{(t)}(r)$. In the distributed version, node i stores the i-th coordinate of each vector, thus requiring $O(r)$ storage capability. The vectors are updated by letting $\widetilde{x}^{(t)}(a) = S^{(t)}x^{(t)}(a)$. and then orthonormalizing $\widetilde{x}^{(t)}(1), \dots \widetilde{x}^{(t)}(r)$ to get $x^{(t)}(1), \dots x^{(t)}(r)$. Orthonormalization can be done locally at each node if it has access to the Gram matrix $G = (G_{ab})_{1\le a,b\le r}$

$$G_{ab} \equiv \frac{1}{n}\sum_{i=1}^n \widetilde{x}_i^{(t)}(a)\widetilde{x}_i^{(t)}(b) \tag{9}$$

This can be computed via gossip averaging, using messages consisting of $r(r+1)/2$ real numbers. Therefore the total communication complexity per node per iteration is of order $r^2\log(1/\varepsilon)$ to achieve precision ε. Indeed, such distributed orthonormalization procedure was studied in [21] for decentralized implementation of the standard power method.

Richer stochastic models for random sparsification. Our main result holds under the assumption that $S^{(1)}$, $S^{(2)}$,...,$S^{(t)}$ are i.i.d. sparsifications of the matrix M. This is a reasonable assumption when the random sparsifications are generated by the algorithm itself. The same assumption can also model short time-scale link failures, as due for instance to fast fading in a wireless setting. On the other hand, a more accurate model of link failures would describe $S^{(1)}$, $S^{(2)}$,...,$S^{(t)}$ as a stochastic process. We think that our main result is generalizable to this setting under appropriate ergodicity assumptions on this process. More explicitly, as long as the underlying stochastic process mixes (i.e. loses memory of its initial state) on time scales shorter than t, the qualitative features of Theorem 2.1 should remain unchanged. Partial support of this intuition is provided by the celebrated Oseledets' multiplicative ergodic theorem that guarantees convergence the exponential growth rate of $\|x^{(t)}\|$ in a very general setting [25] (namely within the context of ergodic dynamical systems).

Communication constraints: Rate and noise. In a decentralized setting, it is unavoidable to take into consideration communication rate constraints and communication errors. The presence of errors implies that the actual matrix used at iteration t is not $S^{(t)}$ but is rather a perturbation of it. The effect of noise can then be studied through Theorem 2.1. Rate constraints imply that real numbers cannot be communicated through the network, unless some quantization is used. An approach consists in using some form of ran-

domized rounding for quantization. In this case, the effect of quantization can also be studied through Theorem 2.1. This implies that, roughly speaking, the error in the eigenvector computed with this approach scales quadratically in the quantization step. (Notice that quantization also affects the vector on the right-hand side of Eq. (2), but we expect this effect to be roughly of the same order as the effect of the quantization of $S^{(t)}$.) Further, when the matrix M itself is sparse, or a fixed sparsification S is used within the ordinary power method, Theorem 2.1 can be used to study the effect of noise and quantization.

4. RELATED WORK

The need for spectral analysis of massive data sets has motivated a considerable effort towards the development of randomized low complexity methods. A short sample of the theoretical literature in this topic includes [9, 4, 10, 14, 12, 11]. Two basic ideas are developed in this line of research: *sparsify* of the original matrix M to reduce the cost of matrix-vector multiplication; *apply* the matrix M to a random set of vectors in order to approximate its range. Both of these approaches are developed in a centralized setting where a single dataset is sent to a central processor. While this allows for more advanced algorithms than power iteration, these algorithms might not be directly applicable in a decentralized setting considered in this paper, where each node has limited computation and comunication capability and the datasets are often extremely large such that the data has to be stored in a distributed manner.

Fast routines for low-rank approximation are useful in many areas of optimization, scientific computing and simulations. Hence similar ideas were developed in that literature: we refer to [16] for references and an overview of the topic.

Kempe and McSherry [21] studied a decentralized power iteration algorithm for spectral analysis. They considered matrices that are inherently sparse. Therefore, no sparsification is used and all the entries are exploited at every iteration. Hence, their algorithm eventually computes the optimal low-rank approximation exactly. The same paper introduced the decentralized orthonormalization mentioned in Section 3.3.

The idea of using a sequence of distinct sparsifications to improve the accuracy of power iteration was not studied in this context. Somewhat related is the basic idea in randomized algorithms for gossip averaging [5]. As discussed in Section 3.2, these algorithms operate by applying a sequence of i.i.d. random matrices to an initial vector of data. The behavior and analysis is however considerably simplified by the fact that these matrices share a common leading eigenvector, that is known *a priori*, namely the eigenvector $u = (1/\sqrt{n}, \dots, 1/\sqrt{n})$. Overviews of this literature is provided by [27] and [8]. Quantization is an important concern in the practical implementation of gossip algorithms, and has been studied in particular in the context of consensus [20, 13]. As discussed in the last section, the effect of randomized quantization can also be included in the present setting.

Finally, there has been recent progress in the development of sparsification schemes that imply better error guarantees than in Eq. (1), see for instance [28]. It would be interesting to study the effect of such sparsification methods in the present setting.

5. PROOF OF THE MAIN THEOREM

In this section, we analyze the quality of the estimation $\widehat{u}^{(t)}$ provided by our algorithm and prove Theorem 2.1. Before diving into the technical argument, it is worth motivating the main ideas. We are interested in analyzing the random trajectory $\{x^{(t)}\}_{t\geq 0}$ defined as per Eq. (2). One difficulty is that this process cannot be asymptotically stationary, since $x^{(t)}$ gets multiplied by a random quantity. Hence it will either grow exponentially fast or shrink exponentially fast.

A natural solution to this problem would be to track the normalized vectors $\widetilde{x}^{(t)} \equiv x^{(t)}/\|x^{(t)}\|$. Also this approach presents some technical difficulty that can be grasped by considering the special case in which $S^{(t)} = M$ for all t (no sparsification is used). Neglecting exceptional initial conditions (such that $\langle x^{(0)}, u\rangle = 0$) this sequence can either converge to u or to $-u$. In particular, it cannot be uniformly convergent. The right way to eliminate this ambiguity is to track the unit vectors $\widetilde{x}^{(t)}$ 'modulo overall sign'. The space of unit vectors modulo a sign is the projective space P_n, that we will introduce more formally below.

We are therefore naturally led to consider the random trajectory $\{\mathbf{x}_t\}_{t\geq 0}$ –indeed a Markov chain– taking values in the projective space $\mathbf{x}_t \in \mathsf{P}_n$. We will prove that two important facts hold under the assumptions of Theorem 2.1: (1) The chain converges quickly to a stationary distribution μ; (2) The distance between the baricenter of μ and u is of order θ^2. Fact (1) implies that $\widehat{u}^{(t)}$, cf. Eq. (4), is a good approximation of the baricenter of μ. Fact (2) then implies Theorem 2.1.

In the next subsection we will first define formally the process $\{\mathbf{x}_t\}_{t\geq 0}$, and provide some background (Section 5.1), and then present the formal proof (Section 5.2), along the lines sketched above.

5.1 The Kesten-Furstenberg Markov chain

As anticipated above, we shall denote by P_n the projective space in $\mathbb{R}^n$. This is defined as the space of lines through the origin in $\mathbb{R}^n$. Equivalently, P_n is the space of equivalence classes in $\mathbb{R}^n \setminus \{0\}$ for the equivalence relation $\sim_{\mathsf{P}}$, such that $x \sim_{\mathsf{P}} y$ if and only if $x = \lambda y$ for some $\lambda \in \mathbb{R} \setminus \{0\}$. This corresponds with the description given above, since it coincides with the space of equivalence classes in $S^n \equiv \{x \in \mathbb{R}^n : \|x\| = 1\}$ for the equivalence relation $\sim_{\mathsf{P}}$, such that $x \sim_{\mathsf{P}} y$ if and only if $x = \lambda y$ for some $\lambda \in \{+1, -1\}$.

In the future, we denote elements of P_n by boldface letters $\mathbf{x}, \mathbf{y}, \mathbf{z}, \dots$ and the corresponding representatives in $\mathbb{R}^n$ by $x, y, z, \dots$. We generally take these representatives to have unit norm. We use a metric on this space defined as

$$d(\mathbf{x}, \mathbf{y}) \equiv \sqrt{1 - \langle x, y\rangle^2}\,.$$

Random elements in P_n will be denoted by boldface capitals $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \dots$.

An invertible matrix $S \in \mathbb{R}^{n\times n}$ acts naturally on P_n, by mapping $\mathbf{x} \in \mathsf{P}_n$ (with representative x) to the element $\mathbf{y} \in$

P_n with representative of Sx (namely the line through Sx, or the unit vector $Sx/\|Sx\|$ modulo sign). We will denote this action by writing $\mathbf{y} = S\mathbf{x}$, but emphasize that it is a non-linear map, since it implicitly involves normalization.

Given a sequence of i.i.d. random matrices $\{S^{(t)}\}_{t\ge 1}$ that are almost surely invertible, with common distribution p_S, we define the Markov chain $\{\mathbf{X}_t\}_{t\ge 0}$ with values in P_n by letting

$$\mathbf{X}_t = S^{(t)}S^{(t-1)}\cdots S^{(1)}\mathbf{X}_0\ , \qquad (10)$$

for all $t \ge 1$. We assume the following conditions:

- L1. There exists no proper linear subspace $V \subseteq \mathbb{R}^n$ such that $S^{(1)}V \subseteq V$ almost surely.
- L2. There exist a sequence $\{S^{(t)}\}_{t\ge 1}$ in the support of p_S, such that letting $S^T \equiv S^{(T)}S^{(T-1)}\cdots S^{(1)}$, we have $\sigma_2(S^T)/\sigma_1(S^T) \to 0$ as $T \to \infty$.

It was proved in [23], that, under the assumptions L1 and L2, there exists a unique measure μ on P_n that is stationary for the Markov chain $\{\mathbf{X}_t\}$. The Markov chain converges to the stationary measure as $t \to \infty$ (we refer to the Appendix for a formal statement).

For the purpose of proving Theorem 2.1, uniqueness of the stationary measure is not enough: we will need to control the rate of convergence to stationarity. We present here a general theorem to bound the rate of convergence, and we will apply it to the chain of interest in the next section. Let us start by stating two more assumptions. We denote by $\mathsf{G} \subseteq \mathsf{P}_n$ a (measurable) subset of the projective space, and assume that there exists a constant $\rho \in (0,1)$ such that

- A1. For any $\mathbf{x} \in \mathsf{G}$, $S^{(t)}\mathbf{x} \in \mathsf{G}$ almost surely.
- A2. For any $\mathbf{x} \neq \mathbf{y} \in \mathsf{G}$, $\mathbb{E}\left[d(S^{(t)}\mathbf{x}, S^{(t)}\mathbf{y})\right] \le \rho\, d(\mathbf{x},\mathbf{y})$.

We then have the following.

THEOREM 5.1. *Assume conditions* L1 *and* L2 *hold, together with* A1 *and* A2. *Denote by μ the unique stationary measure of the Markov chain $\{\mathbf{X}_t\}_{t\ge 0}$. Then*

$$\mu(\mathsf{G}^c) = 0\ .$$

Further, if $\mathbf{X}_0 \in \mathsf{G}$ then for any L-Lipschitz function[1] *$f : \mathsf{P}_n \to \mathbb{R}$, we have*

$$\left|\mathbb{E}[f(\mathbf{X}_t)] - \mu(f)\right| \le\ L\,\rho^t\ .$$

The proof of this Theorem uses a coupling technique analogous to the one of [23]. We present it in the appendix for greater convenience of the reader.

5.2 Proof of Theorem 2.1

In this section we analyze the GOSSIP PCA algorithm using the general methodology developed above. In particular, we consider the Markov chain (10) whereby $\{S^{(\ell)}\}_{1\le\ell\le t}$ are i.i.d. sparsifications of M satisfying the conditions: *(i)* $\|S^{(\ell)} - M\|_2 \le \theta\|M\|_2$; *(ii)* $\mathbb{E}[S^{(\ell)}] = M$; *(iii)* $S^{(\ell)}$ is almost surely non-singular. Throughout the proof, we let $\mathbf{u} \in \mathsf{P}_n$ denote an element of P_n represented by u.

Note that the conditions L1 stated in the previous section holds by assumption in Theorem 2.1. Further let $\lambda_1(\ell)$ and $\lambda_2(\ell)$ the largest and second largest singular values of $S^{(\ell)}$. By assumption *(i)*, and since by hypothesis $\theta \le (1/40)(1-l_2)^{3/2}$, implying $\|S^{(\ell)} - M\|_2 \le (\lambda - |\lambda_2|)/2$, we have $|\lambda_1(\ell)/\lambda_2(\ell)| > 1$ almost surely. Hence by taking $S^{(1)} = S^{(2)} = \cdots = S^{(T)} = \ldots$ in the support of p_S, we have that condition L2 holds as well.

By applying the main theorem in [23] (restated in the Appendix), we conclude that there exists a unique stationary distribution μ for the Markov chain $\{\mathbf{X}_t\}$, and that the chain converges to it.

We next want to apply Theorem 5.1 to bound the support and the rate of convergence to this stationary distribution. We define the 'good' subset $\mathsf{G} \subseteq \mathsf{P}_n$ by

$$\mathsf{G} = \left\{\mathbf{x} \in \mathsf{P}_n : d(\mathbf{x},\mathbf{u}) \le \frac{2\theta}{1-l_2}\right\}\ . \qquad (11)$$

Our next lemma shows assumptions A1 and A2 are satisfied in this set G, with a very explicit expression for the contraction coefficient ρ.

LEMMA 5.2. *Under the hypothesis of Theorem 2.1, for any $\mathbf{x} \in \mathsf{G}$ we have $S^{(\ell)}\mathbf{x} \in \mathsf{G}$. Further, for any $\mathbf{x} \neq \mathbf{y} \in \mathsf{G}$, letting $\rho \equiv 1 - (4/5)(1-l_2) \in (0,1)$, we have*

$$\mathbb{E}d(S^{(\ell)}\mathbf{x}, S^{(\ell)}\mathbf{y}) \le \rho\, d(\mathbf{x},\mathbf{y})\ .$$

The proof of this lemma can be found in the next section. As a consequence of this lemma we can apply Theorem 5.1. In particular, we conclude that μ is supported on the good set G.

Next consider the estimate $\widehat{u}^{(t)} \in \mathbb{R}^n$ produced by our algorithm, cf. Eq. (4). This is given in terms of the Markov chain on P_n by

$$\widehat{u}^{(t)} = \frac{\sum_{\ell=1}^t f(\mathbf{X}_\ell)}{\|\sum_{\ell=1}^t f(\mathbf{X}_\ell)\|}\ ,$$

where we define $f : \mathsf{P}_n \mapsto \mathbb{R}^n$ such that $f(\mathbf{x})$ is a representative of $\mathbf{x}$ satisfying $\|f(\mathbf{x})\| = 1$ and $\langle u, f(\mathbf{x})\rangle \ge 0$. We use $\mathbf{U}_t \in \mathsf{P}_n$ to denote an element in P_n represented by $\widehat{u}^{(t)}$.

Let $\mu(f) = \int f(\mathbf{x})\mu(\mathrm{d}\mathbf{x}) \in \mathbb{R}^n$ be the expectation of $f(\,\cdot\,)$ with respect to the stationary distribution (informally, this is the baricenter of μ). With a slight abuse of notation, we let $\mu(f)$ denote the corresponding element in P_n as well. Then, by the triangular inequality, we have, for any t,

$$d(\mathbf{u},\mathbf{U}_t) \le d(\mathbf{u},\mu(f)) + d(\mathbf{U}_t,\mu(f))\ .$$

The left hand side is the error of our estimate of the leading eigenvector. This is decomposed in two contributions: a deterministic one, namely $d(\mathbf{u},\mu(f))$, that gives the distance between the leading eigenvector and the baricenter of μ, and a random one i.e. $d(\mathbf{U}_t,\mu(f))$, that measures the distance between the average of our sample and the average of the distribution.

[1] We say that f is L-Lipschitz if, for any $\mathbf{x},\mathbf{y} \in \mathsf{P}_n$, $|f(\mathbf{x}) - f(\mathbf{y})| \le L\, d(\mathbf{x},\mathbf{y})$.

In order to bound $d(\mathbf{U}_t, \mu(f))$, we use the following fact that holds for any $a, b \in \mathbb{R}^n$

$$\sqrt{1 - \frac{\langle a,b\rangle^2}{\|a\|^2\|b\|^2}} \leq \frac{\|a-b\|}{\sqrt{\|a\|\|b\|}} \,. \tag{12}$$

This follows immediately from $2\|a\|\|b\| - 2\langle a,b\rangle \leq \|a-b\|^2$. We apply this inequality to $a = \mu(f)$ and $b = (1/t)\sum_{\ell=1}^t f(\mathbf{X}_\ell)$. We need therefore to lower bound $\|\mu(f)\|$ and $\|(1/t)\sum_{\ell=1}^t f(\mathbf{X}_\ell)\|$ and to upper bound $\|(1/t)\sum_{\ell=1}^t f(\mathbf{X}_\ell) - \mu(f)\|$.

Denote by $\mathcal{P}_u$ the orthogonal projector onto u. From Theorem 5.1, we know that $\mu(\mathsf{G}^c) = 0$. Hence, using $\theta < (1/40)(1-l_2)^{3/2}$, we have $\|\mu(f)\| \geq \|\mu(\mathcal{P}_u(f))\| \geq \sqrt{1-1/400}$. Similarly since $\mathbf{X}_0 \in \mathsf{G}$, we have by A1 that $\mathbf{X}_\ell \in \mathsf{G}$ for all ℓ, and therefore $\|(1/t)\sum_{\ell=1}^t f(\mathbf{X}_\ell)\| \geq \sqrt{1-1/400}$. We are left with the task of bounding $\|a-b\|$. This is done in the next lemma that uses in a crucial way Theorem 5.1.

LEMMA 5.3. *Under the hypothesis of Theorem 2.1*

$$\mathbb{E}\Big\|\frac{1}{t}\sum_{\ell=1}^t f(\mathbf{X}_\ell) - \mu(f)\Big\|^2 \leq \frac{70\theta^2}{(1-l_2)^2 t} \,.$$

Applying Markov's inequality and Eq. (12), we get, with probability larger than $1-\delta/2$

$$d(\mathbf{U}_t, \mu(f)) \leq \frac{12\theta}{(1-l_2)\sqrt{t\delta}} \,.$$

Next, we bound the term $d(\mathbf{u}, \mu(f))$ in Eq. (12) with the following lemma.

LEMMA 5.4. *Under the hypothesis of Theorem 2.1,*

$$d(\mathbf{u}, \mu(f)) \leq 8\Big(\frac{\theta\log(1/\theta)}{(1-l_2)}\Big)^2 \,.$$

By noting that $\|u - \widehat{u}^{(t)}\| \leq \sqrt{2}\, d(\mathbf{u}, \widehat{\mathbf{U}}_t)$, this finishes the proof of the theorem.

6. PROOF OF TECHNICAL LEMMAS

6.1 Proof of Remark 2.2

Assuming initial vector $X \in \mathbb{R}^n$ with i.i.d. Gaussian entries, we can get close to u by iteratively applying a single sparsification S. Define a good set of initial vectors

$$\mathcal{F}_n = \left\{ x \in \mathbb{R}^n : |u^* x| \geq \frac{1}{n^{5/2}} \text{ and } \max_{i\in[n]} |u_i^* x| \leq \sqrt{\frac{6\log n}{n}} \right\} \,.$$

Since, $u_i^* X$'s are independent and distributed as $\mathsf{N}(0, 1/n)$, it follows that we have $\mathbb{P}(|u_i^* X| \geq \sqrt{(6\log n)/n}) \leq 2/n^3$ and $\mathbb{P}(|u^* X|) \leq 1/n^2$. Applying union bound, we get $\mathbb{P}(X \in \mathcal{F}_n) \geq 1 - 3/n^2$. Assuming we start from this good set, we show that for k large enough, we are guaranteed to have $\|u - \mathbf{x}^{(k)}\| \leq \theta/(1-l_2)$.

Let $\{\widetilde{\lambda}_i\}$ be the eigenvalues of S such that $\widetilde{\lambda}_1 \geq |\widetilde{\lambda}_2| \geq \cdots \geq |\widetilde{\lambda}_n|$, and let $\{\widetilde{u}_i\}$ be the corresponding eigenvectors. We know that $\widetilde{\lambda}_1 > 0$ since $\widetilde{\lambda}_1 \geq \lambda - \|S-M\|_2$ and $\|S-M\|_2 < \lambda$ by assumption. Then, by the triangular inequality,

$$\|u - \mathbf{x}^{(k)}\| \leq \|u - \widetilde{u}\| + \|\widetilde{u} - \mathbf{x}^{(k)}\| \,.$$

To bound the first term, note that

$$\begin{aligned} \|M-S\|_2 &\geq |u^t(M-S)u| \\ &\geq \lambda - \widetilde{\lambda}_1 (u^*\widetilde{u})^2 - \widetilde{\lambda}_2 \|\mathcal{P}_{\widetilde{u}^\perp}(u)\|^2 \,. \end{aligned}$$

This implies that $(u^*\widetilde{u})^2 \geq (\lambda - \widetilde{\lambda}_2 - \|M-S\|_2)/(\widetilde{\lambda}_1 - \widetilde{\lambda}_2)$. We can further apply Weyl's inequality [18], to get $|\widetilde{\lambda}_i - \lambda_i| \leq \|M-S\|_2$. It follows that $(u^*\widetilde{u})^2 \geq (\lambda - \lambda_2 - 2\|M-S\|_2)/(\lambda - \lambda_2 + 2\|M-S\|_2)$. Note that this bound is non-trivial only if $\|M-S\|_2 \leq (\lambda-\lambda_2)/2$. Using the fact that $(1-a)/(1+a) \leq (1-a)^2$ for any $|a| < 1$, this implies that

$$\|u - \widetilde{u}\| \leq \sqrt{\frac{4\|M-S\|_2}{\lambda - \lambda_2}} \,.$$

In particular, for $\|M-S\|_2 \leq \theta^2\|M\|_2/(2(1-l_2))$ as per our assumption, this is less than $\sqrt{2}\theta/(1-l_2)$.

To bound the second term, we use $x^{(0)} \in \mathcal{F}_n$ to get

$$\begin{aligned} \frac{(\widetilde{u}^* S^k x^{(0)})^2}{\|S^k x^{(0)}\|^2} &\geq \frac{1}{1 + \sum_{i\geq 2} \frac{\widetilde{\lambda}_i^{2k}}{\widetilde{\lambda}_1^{2k}} \frac{(\widetilde{u}_i^* x^{(0)})^2}{(\widetilde{u}_1^* x^{(0)})^2}} \\ &\geq 1 - (\widetilde{\lambda}_2/\widetilde{\lambda}_1)^{2k} 6n^5 \log n \\ &\geq 1 - \frac{\theta^2}{4(1-l_2)^2} \,. \end{aligned}$$

In the last inequality we used $k \geq 3\log(n/\theta)/(1-l_2-\theta)$, and the fact that $(\widetilde{\lambda}_2/\widetilde{\lambda}_1) \leq l_2 + \theta$. Then, $\|\widetilde{u} - \mathbf{x}^{(k)}\| \leq \theta/(\sqrt{2}(1-l_2))$. Collecting both terms, this proves the desired claim. □

6.2 Proof of Lemma 5.2

LEMMA 6.1 (CONTRACTION). *For a given $\nu \leq (1/20)$, assume that x, x' satisfy $\|x\| = \|x'\| = 1$, $\langle u, x\rangle \geq 0$, $\langle u, x'\rangle \geq 0$, $\|\mathcal{P}_{u_\perp}(x)\| \leq \nu$, and $\|\mathcal{P}_{u_\perp}(x')\| \leq \nu$. Then, under the hypothesis of Lemma 5.2, we have*

$$\left\|\mathcal{P}_{u_\perp}\left(\frac{Qx}{\|Qx\|} - \frac{Qx'}{\|Qx'\|}\right)\right\| \leq \big(l_2(1+3\nu^2) + 3\theta\big)\|z - z'\| \,, \tag{13}$$

and

$$\left\|\mathcal{P}_{u}\left(\frac{Qx}{\|Qx\|} - \frac{Qx'}{\|Qx'\|}\right)\right\| \leq \big(4\nu + 4\theta\big)\|z - z'\| \,, \tag{14}$$

where $l_2 \equiv |\lambda_2|/\lambda$, $z = \mathcal{P}_{u_\perp}(x)$ and $z' = \mathcal{P}_{u_\perp}(x')$.

PROOF. By the assumption that $\langle u, x\rangle \geq 0$ and $\langle u, x'\rangle \geq 0$, we have $x = \sqrt{1-\|z\|^2}u + z$ and $x' = \sqrt{1-\|z'\|^2}u + z'$. The following inequalities, which follow from $\|Q - M\|_2 \leq \theta\lambda$, will be frequently used.

$$\begin{aligned} (1-\theta)\lambda \;\leq\; \|Qu\| \;\leq\; (1+\theta)\lambda \,, \\ (l_2-\theta)\lambda \;\leq\; \|Qz\| \;\leq\; (l_2+\theta)\lambda \,. \end{aligned}$$

The following inequalities will also be useful in the proof.

$$\|x - x'\| \leq (1/\sqrt{1-\nu^2})\|z - z'\| \,, \tag{15}$$

where we used $\sqrt{1-a^2}-\sqrt{1-b^2} \le (\nu/\sqrt{1-\nu^2})|a-b|$ for $|a| \le \nu$ and $|b| \le \nu$. Similarly, using the fact that Mu and Mz are orthogonal

$$\begin{aligned}\|Qx\| &\ge (\sqrt{1-\|z\|^2})\|Mu\| - \|Q-M\|_2 \\ &\ge \lambda(\sqrt{1-\nu^2}-\theta)\,, \end{aligned} \tag{16}$$

Next, we want to show that

$$\left|\frac{1}{\|Qx\|}-\frac{1}{\|Qx'\|}\right| \le \frac{(2.2\nu+0.1\theta)}{(\sqrt{1-\nu^2}-\theta)^3}\|z-z'\|\,. \tag{17}$$

We use the equality $1/a-1/b=(a^2-b^2)/(ab(a+b))$ with $a=\|Qx\|$ and $b=\|Qx'\|$. The denominator can be bounded using (16). It is enough to bound $|\|Qx'\|^2-\|Qx\|^2|$ using

$$\begin{aligned}&\left|\|Q(\sqrt{1-\|z\|^2}u+z)\|^2-\|Q(\sqrt{1-\|z'\|^2}u+z')\|^2\right| \\ &\le \left|\|z\|^2-\|z'\|^2\right|\|Qu\|^2+\left|\|Qz\|^2-\|Qz'\|^2\right| \\ &\quad +2\left|(\sqrt{1-\|z\|^2}z-\sqrt{1-\|z'\|^2}z')^*Q^*Qu\right|\,. \end{aligned}$$

Note that $\left|\|z\|^2-\|z'\|^2\right|\|Qu\|^2 \le 2\nu(1+\theta)^2\lambda^2\|z-z'\|$, and $\left|\|Qz\|^2-\|Qz'\|^2\right| \le 2\nu(l_2+\theta)^2\lambda^2\|z-z'\|$. The last term can be decomposed into

$$\begin{aligned}&2\left|(\sqrt{1-\|z\|^2}z-\sqrt{1-\|z'\|^2}z')^*Q^*Qu\right| \\ &\le 2|\sqrt{1-\|z\|^2}-\sqrt{1-\|z'\|^2}|\,|z^*Q^*Qu| \\ &\quad +2\sqrt{1-\|z'\|^2}\,|(z-z')^*Q^*Qu|\,. \end{aligned}$$

Note that $|\sqrt{1-\|z\|^2}-\sqrt{1-\|z'\|^2}| \le (\nu/\sqrt{1-\nu^2})\|z-z'\|$, $|z^*Q^*Qu| \le \lambda^2\theta(l_2+\theta)$, and $|(z-z')^*Q^*Qu| \le \lambda^2(l_2+\theta)\theta\|z-z'\|$. Collecting all the terms and assuming $\theta \le 1/40$ and $\nu \le 1/20$, $|\|Qx\|-\|Qx'\|| \le (4.4\nu+0.1\theta)\lambda^2\|z-z'\|$. this implies (17).

To prove (13), define $T_1 \equiv \mathcal{P}_{u^\perp}(Qx-Qx')/\|Qx\|$ and $T_2 \equiv \mathcal{P}_{u^\perp}(Qx')\big((1/\|Qx\|)-(1/\|Qx'\|)\big)$. We bound each of these separately.

$$\begin{aligned}\|T_1\| &= \frac{\left\|\mathcal{P}_{u^\perp}\big(M(x-x')+(Q-M)(x-x')\big)\right\|}{\|Qx\|} \\ &\overset{(a)}{\le} \frac{l_2\|z-z'\|+\theta\|x-x'\|}{(\sqrt{1-\nu^2}-\theta)} \\ &\overset{(b)}{\le} \frac{l_2+(\theta/\sqrt{1-\nu^2})}{(\sqrt{1-\nu^2}-\theta)}\|z-z'\|, \end{aligned}$$

where (a) follows from (16) and the fact that $\mathcal{P}_{u^\perp}Mu=0$, and (b) follows from (15). Similarly, using (17)

$$\begin{aligned}\|T_2\| &= \left\|\mathcal{P}_{u^\perp}\big(Q(\sqrt{1-\|z'\|^2}u+z')\big)\right\|\left|\frac{1}{\|Qx\|}-\frac{1}{\|Qx'\|}\right| \\ &\le (\theta+\nu l_2)\frac{2.2\nu+0.1\theta}{(\sqrt{1-\nu^2}-\theta)^3}\|z-z'\|\,. \end{aligned}$$

Notice that by assumption, we have $\theta \le (1/40)$, and by the definition of G in (11), we have $\nu \le (1/20)$. Then, after some calculations, we have proved (13). Analogously we can prove (14) by bounding $T_3 \equiv \mathcal{P}_u(Qx-Qx')/\|Qx\|$ and $T_4 \equiv \mathcal{P}_u(Qx')\big((1/\|Qx\|)-(1/\|Qx'\|)\big)$ separately. □

We are now in position to prove Lemma 5.2.

Proof of Lemma 5.2. We first show that for any $\mathbf{x} \in \mathsf{G}$ with a representative x such that $\langle x,u\rangle \ge 0$, we have $Q\mathbf{x} \in \mathsf{G}$. Note that, by triangular inequality, $\|\mathcal{P}_{u^\perp}(Qu)\| \le \theta\lambda$ and $\|Qu\| \ge (1-\theta)\lambda$. Applying Lemma 6.1 to x and u, we get

$$\begin{aligned}&\left\|\mathcal{P}_{u_\perp}\left(\frac{Qx}{\|Qx\|}\right)\right\| \\ &\le \left\|\mathcal{P}_{u_\perp}\left(\frac{Qu}{\|Qu\|}\right)\right\|+\Big(l_2\Big(1+3\nu^2\Big)+3\theta\Big)\,\|\mathcal{P}_{u_\perp}(x-u)\| \\ &\le \Big(\frac{1}{1-\theta}+3\nu\Big)\theta+\Big(l_2\Big(1+3\nu^2\Big)\Big)\nu\,. \end{aligned} \tag{18}$$

For $\theta \le (1/40)$ and for θ and ν satisfying,

$$\frac{2}{1-l_2}\theta \le \nu \le \min\left\{\sqrt{\frac{2(1-l_2)}{15}}\,,\ \frac{1}{20}\right\},$$

the right-hand side of (18) is always smaller than ν, since $\big((1/(1-\theta))+3\nu\big)\theta \le (3/5)(1-l_2)\nu$ and $3\nu^2 \le (2/5)(1-l_2)$. This proves our claim for $\theta \le (1/40)(1-l_2)^{3/2}$ and $\nu \in [(2\theta)/(1-l_2), \sqrt{1-l_2}/20]$ as per our assumptions.

Next, we show that there is a contraction in the set G. For x and x' satisfying the assumptions in Lemma 6.1, define $y \equiv Ax/\|Ax\|$, $y' \equiv Ax'/\|Ax'\|$, $z \equiv \mathcal{P}_{u^\perp}(x)$, and $z' \equiv \mathcal{P}_{u^\perp}(x')$. For $\|x\| \le \nu$ and $\|x'\| \le \nu$ we have

$$1-2\nu^2 \le (x,x') \le 1.$$

Using the above bounds we get

$$\frac{1-\langle y,y'\rangle^2}{1-\langle x,x'\rangle^2} \le \frac{1}{(1-\nu^2)}\frac{1-\langle y,y'\rangle}{1-\langle x,x'\rangle}\,.$$

We can further bound $(1-\langle y,y'\rangle)/(1-\langle x,x'\rangle)$ using Lemma 6.1.

$$\|y-y'\| \le \sqrt{(l_2+3\nu^2+3\theta)^2+(4\nu+4\theta)^2}\,\|z-z'\|\,.$$

Using $\|z-z'\|^2 \le \|x-x'\|^2 = 2-2\langle x,x'\rangle$, we get

$$\frac{1-\langle y,y'\rangle}{1-\langle x,x'\rangle} \le \big(l_2+3\nu^2+3\theta\big)^2+(4\nu+4\theta)^2\,.$$

For $\theta \le (1/40)(1-l_2)^{3/2}$ and $\nu \le \sqrt{1-l_2}/20$ as per our assumptions, it follows, after some algebra, that

$$\sqrt{\frac{1-\langle y,y'\rangle^2}{1-\langle x,x'\rangle^2}} \le \sqrt{\frac{(l_2+3\nu^2+3\theta)^2+(4\nu+4\theta)^2}{1-\nu^2}} \le \rho\,,$$

for $\rho \ge 1-0.8(1-l_2)$.

□

6.3 Proof of Lemma 5.3

Expanding the summation, we get

$$\begin{aligned}&\left\|\frac{1}{t}\sum_{s=1}^{t}f(\mathbf{X}_s)-\mu(f)\right\|^2 \\ &\quad=\frac{1}{t^2}\sum_{s=1}^{t}\|f(\mathbf{X}_s)-\mu(f)\|^2+ \\ &\quad\frac{2}{t^2}\sum_{r=1}^{t}\sum_{r<s}\big\langle f(\mathbf{X}_r)-\mu(f),f(\mathbf{X}_s)-\mu(f)\big\rangle\,, \end{aligned}$$

where $\langle a, b\rangle = a^*b$ denotes the scalar product of two vectors. We can bound the first term by $20\theta^2/(t(1-l_2)^2)$, since

$$\|f(\mathbf{X}_s) - \mu(f)\|^2 \le \frac{20\theta^2}{(1-l_2)^2}\ , \tag{19}$$

where we used $\|\mathcal{P}_u(f(\mathbf{X}_s) - \mu(f))\|^2 \le 4\theta^2/(1-l_2)^2$ and $\|\mathcal{P}_{u^\perp}(f(\mathbf{X}_s) - \mu(f))\|^2 \le 16\theta^2/(1-l_2)^2$ for $\mathbf{X}_s \in \mathsf{G}$.

To bound the second term, let $y \equiv f(\mathbf{X}_r) - \mu(f)$. Note that by (19), $\|y\| \le \sqrt{20}\theta/(1-l_2)$. We apply Theorem A.3 together with Lemma 5.2 to get

$$\Big|\mathbb{E}\big[y^* f(\mathbf{X}_s) \,\big|\, \mathbf{X}_r\big] - y^*\mu(f)\Big| \le \rho^{s-r}\|y\|^2\ ,$$

for $r < s$. Using the fact that for $|\rho| \le 1$, $\sum_{r=1}^t \sum_{r<s} \rho^{s-r} \le \sum_{r=1}^t \rho^{-r}\rho^{r+1}/(1-\rho) \le \rho t/(1-\rho)$, it follows that

$$\begin{aligned}
&\sum_{r=1}^{t}\sum_{r<s}\mathbb{E}\Big[\big\langle f(\mathbf{X}_r) - \mu(f), f(\mathbf{X}_s) - \mu(f)\big\rangle\Big] \\
&\quad\le \sum_{r=1}^{t}\sum_{r<s}\mathbb{E}\Big[\big\langle f(\mathbf{X}_r) - \mu(f), \mathbb{E}\big[f(\mathbf{X}_s) - \mu(f) \,\big|\, \mathbf{X}_r\big]\big\rangle\Big] \\
&\quad\le \frac{20\theta^2}{(1-l_2)}\sum_{r=1}^{t}\sum_{r<s}\rho^{s-r} \le \frac{20\theta^2\rho t}{(1-l_2)(1-\rho)}\ .
\end{aligned}$$

Combining the above bounds we get

$$\mathbb{E}\Big\|\frac{1}{t}\sum_{s=1}^{t} f(\mathbf{X}_s) - \mu(f)\Big\|^2 \le \frac{20\theta^2}{t(1-l_2)^2} + \frac{40\theta^2\rho}{(1-l_2)(1-\rho)t}.$$

For $\rho = 1 - (4/5)(1-l_2)$ as in Lemma 5.2, this proves the desired claim. □

6.4 Proof of Lemma 5.4

From Theorem 5.1, we know $\mu(\mathsf{G}^c) = 0$. This imples that $\|\mu(f)\|^2 \ge \|\mathcal{P}_u(\mu(f))\|^2 \ge 1 - 1/400$. Then,

$$d(\mathbf{u}, \mu(f)) = \frac{\|\mathcal{P}_{u^\perp}(\mu(f))\|}{\|\mu(f)\|} \le 2\|\mathcal{P}_{u^\perp}(\mu(f))\|\ .$$

Let $\mathbf{X}$ be a random element in P_n following the stationary distribution $\mu(\cdot)$, and the random vector $X \in \mathbb{R}^n$ be the representative. From the definition of $f(\cdot)$, $\mathcal{P}_{u^\perp}(X)$ is invariant when we apply $f(\cdot)$, whence $\mathcal{P}_{u_\perp}(f(\mathbf{X})) = \mathcal{P}_{u_\perp}(X)$. We can bound $\|\mathcal{P}_{u_\perp}(X)\|$ with the following recursion.

$$\begin{aligned}
\mathcal{P}_{u_\perp}(X) &= \mathbb{E}\left[\frac{\mathcal{P}_{u_\perp}(\prod_{\ell=1}^k S^{(\ell)}X)}{\|\prod_{\ell=1}^k S^{(\ell)}X\|}\,\Big|\,X\right] \\
&= \mathbb{E}\left[\mathcal{P}_{u_\perp}\Big(\prod_{\ell=1}^k S^{(\ell)}X\Big)\Big(\frac{1}{\|\prod_{\ell=1}^k S^{(\ell)}X\|} - \frac{1}{\|M^kX\|}\Big)\,\Big|\,X\right] \\
&\quad + \frac{\mathbb{E}\Big[\mathcal{P}_{u_\perp}\Big(\prod_{\ell=1}^k S^{(\ell)}X\Big)\,\Big|\,X\Big]}{\|M^kX\|} \\
&= \mathbb{E}\Big[\mathcal{P}_{u_\perp}\Big(\prod_{\ell=1}^k S^{(\ell)}X\Big)\Big(\frac{1}{\|\prod_{\ell=1}^k S^{(\ell)}X\|} - \frac{1}{\|M^kX\|}\Big)\,\Big|\,X\Big] \\
&\quad + \frac{\mathcal{P}_{u_\perp}(M^kX)}{\|M^kX\|}\ .
\end{aligned}$$

Let $\nu \equiv 2\theta/(1-l_2)$. To bound the second term, note that $\mu(\mathsf{G}^c) = 0$. This imples that $\|X\| \ge \|\mathcal{P}_u(X)\| \ge \sqrt{1-\nu^2}$ and $\mathcal{P}_{u^\perp}(X) \le \nu$ with probability one. Then,

$$\frac{\big\|\mathcal{P}_{u_\perp}(M^kX)\big\|}{\|M^kX\|} \le l_2^k\frac{\nu}{\sqrt{1-\nu^2}},$$

with probability one. To bound the first term, we use the telescoping sum

$$\prod_{\ell=1}^k S^{(\ell)} - M^k = \sum_{i=1}^k\Big(\prod_{\ell=i+1}^k S^{(\ell)}\Big)(S^{(i)} - M)M^{i-1}\ .$$

Applying the triangular inequality of the operator norm, we have

$$\Big\|\prod_{\ell=1}^k S^{(\ell)} - M^k\Big\|_2 \le \lambda^k\big((1+\theta)^k - 1\big)\ ,$$

which follows from $\Big(\prod_{\ell=i+1}^k S^{(\ell)}\Big)(M^{(i)} - M)M^{i-1} \le \lambda^k(1+\theta)^{k-i}\theta$. Using the above inequality, we get the following bounds with probability one.

$$\begin{aligned}
\Big\|\mathcal{P}_{u_\perp}\Big(\prod_{\ell=1}^k S^{(\ell)}X\Big)\Big\| &\le \|\mathcal{P}_{u_\perp}(M^kX)\| + \Big\|(\prod_{\ell=1}^k S^{(\ell)} - M^k)X\Big\| \\
&\le \lambda^k(l_2^k\nu + (1+\theta)^k - 1)\ ,\ \text{and} \\
\Big\|\mathcal{P}_{u}\Big(\prod_{\ell=1}^k S^{(\ell)}X\Big)\Big\| &\ge \|\mathcal{P}_{u}(M^kX)\| - \Big\|(\prod_{\ell=1}^k S^{(\ell)} - M^k)X\Big\| \\
&\ge \lambda^k(\sqrt{1-\nu^2} - (1+\theta)^k + 1)\ .
\end{aligned}$$

Then it follows that,

$$\begin{aligned}
\left|\frac{1}{\|\prod_{\ell=1}^k S^{(\ell)}X\|} - \frac{1}{\|M^kX\|}\right| &\le \frac{\|(M^k - \prod_{\ell=1}^k S^{(\ell)})X\|}{\|M^kX\|\|\prod_{\ell=1}^k S^{(\ell)}X\|} \\
&\le \frac{((1+\theta)^k - 1)}{\lambda^k\sqrt{1-\nu^2}(\sqrt{1-\nu^2} - (1+\theta)^k + 1)}.
\end{aligned}$$

Collecting all the terms, we get

$$d(\mathbf{u}, \mu(f)) \le \frac{2((1+\theta)^k - 1 + l_2^k\nu)((1+\theta)^k - 1)}{\sqrt{1-\nu^2}(\sqrt{1-\nu^2} - (1+\theta)^k + 1)} + \frac{2l_2^k\nu}{\sqrt{1-\nu^2}}\ .$$

Let $k = \lceil \log(\theta)/\log(l_2)\rceil$ such that $l_2^k \le \theta$. From the assumption that $\theta \le (1-l_2)^{3/2}/40$, it follows that $\theta\log\theta \le 0.12(1-l_2)$. Then,

$$\begin{aligned}
(1+\theta)^{(\log\theta/\log l_2)} - 1 &\le e^{(\theta\log\theta/\log l_2)} - 1 \\
&\le \frac{1.1}{(1-l_2)}\theta\log(1/\theta)\ ,
\end{aligned}$$

Then, after some algebra, $(1+\theta)^k - 1 \le \theta + (1+\theta)((1+\theta)^{(\log\theta/\log l_2)} - 1) \le 1.5\theta\log(1/\theta)/(1-l_2)$, and $(1+\theta)^k - 1 + l_2^k\nu \le 1.5\theta\log(1/\theta)/(1-l_2)$. It also follows that $\sqrt{1-\nu^2} \ge \sqrt{399/400}$ and $(\sqrt{1-\nu^2} - (1+\theta)^k + 1) \ge 0.8$. Collecting all the terms, we get the desired bound on $d(\mathbf{u}, \mu(f))$.

□

7. EIGENVALUE ESTIMATION

In the previous sections, we discussed the challenging task of computing the largest eigenvector under the gossip setting. A closely related task of computing the largest eigenvalue is also practically important in many computational problems. For example, positioning from pairwise distances requires the leading eigenvalues, as well as the leading eigenvectors, to correctly find the positions [24]. In the following, we present an algorithm to estimate the leading eigenvalue under the gossip setting and provide a performance guarantee. Although the proposed algorithm uses the same trajectory $\{x^{(t)}\}$ from GOSSIP PCA, the analysis is completely different from that of the eigenvector estimator.

We assume to have at our disposal the random trajectory $\{x^{(t)}\}_{t\geq 0}$, defined as in (2), possibly from running GOSSIP PCA. Assume that we start with $x^{(0)}$ with entries distributed as $\mathsf{N}(0,1)$. Our estimate for the top eigenvalue λ after t iterations is

$$\widehat{\lambda}^{(t)} = \Big\{ \big| \langle x^{(0)}, x^{(t)} \rangle \big| \Big\}^{1/t} .$$

Although, this estimator uses the same trajectory $\{x^{(t)}\}_{t\geq 0}$ as GOSSIP PCA, the analysis significantly differs from that of the eigenvector estimator. Hence, the statement of the error bound in Theorem 7.1 is also significantly different from Theorem 2.1. The main idea of our analysis is to bound the second moment of the estimate and apply Chebyshev's inequality. Therefore, the second moment of $S_{ij}^{(\ell)}$ characterized by α determines the accuracy of the sparsification.

$$\max_{i,j} \mathrm{Var}(S_{ij}^{(\ell)}) \leq (\alpha/n)\|M\|_2^2 . \qquad (20)$$

The trade-off between d, which determines the complexity, and α depends on the specific sparsification method. With random sampling described in Section 3, it is not difficult to show that Eq. (20) holds with only $\alpha = O(1/d)$.

Our main result bounds the error of the algorithm in terms of α, t, and $\gamma \equiv \sum_{i=1}^n (|\lambda_i|/\lambda)$. The proof of this theorem is outline in the following section.

THEOREM 7.1. *Let $\{S^{(\ell)}\}_{\ell\geq 1}$ be a sequence of i.i.d. $n\times n$ random matrices satisfying $E[S^{(\ell)}] = M$ and Eq. (20). Assume $\alpha < 1/2$ and $\max\{\log_2 n, 2\log_{(1/l_2)} n\} \leq t \leq n/(4\alpha\gamma)$. Then with probability larger than $1 - \max\{\delta, 16/n^2\}$*

$$\left| \frac{\widehat{\lambda}^{(t)} - \lambda}{\lambda} \right| \leq \max\left\{ \frac{8\sqrt{2}}{tn\sqrt{\delta}};\ 32\sqrt{\frac{\alpha\gamma^{3/2}\log n}{t^2\delta}};\ 48\sqrt{\frac{\alpha\gamma^3(\log n)^2}{tn\delta}} \right\} ,$$

provided the right hand side is smaller than $1/t$.

7.1 Proof of Theorem 7.1

The proof idea is fairly simple. Let $x_0 = x^{(0)}$ and define $\lambda^{(t)} \equiv x_0^* x^{(t)}$ such that our estimator is $\widehat{\lambda}^{(t)} \equiv (|\lambda^{(t)}|)^{1/t}$. We will show that this is close to the desired result λ by applying Chebyshev inequality to $\lambda^{(t)}$. In order to do this we need to compute its mean and variance.

LEMMA 7.2. *Consider the two operators $\mathcal{A}, \mathcal{B} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, defined as follows*

$$\mathcal{A}(X) \equiv MXM^* , \qquad (21)$$

$$\mathcal{B}(X) \equiv \frac{\lambda^2\alpha}{n}\langle X, \mathbb{I}_n\rangle \mathbb{I}_n , \qquad (22)$$

where $\langle X, Y\rangle = \mathrm{Tr}(X^*Y)$. Then, conditional on x_0 we have

$$\mathbb{E}[\lambda^{(t)}|x_0] = x_0^* M^t x_0 ,$$
$$Var(\lambda^{(t)}|x_0) \leq \langle x_0x_0^*, (\mathcal{A}+\mathcal{B})^t(x_0x_0^*)\rangle - \langle x_0x_0^*, \mathcal{A}^t(x_0x_0^*)\rangle .$$

The next lemma provides a bound on the variance.

LEMMA 7.3. *Let $\mathcal{A}, \mathcal{B} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ be defined as in Eqs. (21) and (22). Further assume $\alpha < 1/2$ and $\alpha t\gamma < n/4$. Then, for any two vectors x, $y \in \mathbb{R}^n$, $\|x\| = \|y\| = 1$,*

$$\begin{aligned} &\big|\langle yy^*, (\mathcal{A}+\mathcal{B})^t(xx^*)\rangle - \langle yy^*, \mathcal{A}^t(xx^*)\rangle\big| \\ &\leq 4\lambda^{2t}\Big\{ \frac{n\alpha^t + 8\alpha^2\gamma}{4n^2} + \frac{\alpha\sqrt{\gamma}}{n}\Big(\sum_{i=1}^n \frac{|\lambda_i|}{\lambda}\big((u_i^*x)^2 + (u_i^*y)^2\big)\Big) \\ &\quad + \frac{\alpha t\gamma}{n}\Big(\sum_{i=1}^n \frac{|\lambda_i|}{\lambda}(u_i^*x)^2\Big)\Big(\sum_{i=1}^n \frac{|\lambda_i|}{\lambda}(u_i^*y)^2\Big)\Big\} . \end{aligned}$$

For the proof of Lemmas 7.2 and 7.3 we refer to the longer version of this paper. Let $\mathcal{G}_n$ be any measurable subset of $\mathbb{R}^n$. This forms a set of 'good' initial condition x_0, and its complement will be denoted by $\overline{\mathcal{G}}_n$. With an abuse of notation, $\mathcal{G}_n$ will also denote the event $x_0 \in \mathcal{G}_n$ (and analogously for $\overline{\mathcal{G}}_n$). Also let, $\widehat{\lambda} = \widehat{\lambda}^{(t)}$. Then, for any $\Delta > 0$,

$$\begin{aligned} &\mathbb{P}\big\{\widehat{\lambda} \notin [\lambda(1-\Delta)^{1/t}, \lambda(1+\Delta)^{1/t}]\big\} \\ &\leq \mathbb{P}\big\{\big|\widehat{\lambda}^t - \lambda^t\big| \geq \Delta\lambda^t \,\big|\, \mathcal{G}_n\big\} + \mathbb{P}\big\{\overline{\mathcal{G}}_n\big\} \\ &\leq \frac{1}{\Delta^2\lambda^{2t}}\mathbb{E}\big\{(\widehat{\lambda}^t - \lambda^t)^2\big|\,\mathcal{G}_n\big\} + \mathbb{P}\big\{\overline{\mathcal{G}}_n\big\} \\ &\leq \frac{1}{\Delta^2\lambda^{2t}}\big(\mathbb{E}\{\widehat{\lambda}^t|\,\mathcal{G}_n\} - \lambda^t\big)^2 + \frac{1}{\Delta^2\lambda^{2t}}\sup_{x_0\in\mathcal{G}_n}\mathrm{Var}\big(\widehat{\lambda}^t\big|\,x_0\big) + \mathbb{P}\big\{\overline{\mathcal{G}}_n\big\} . \end{aligned}$$

We shall upper bound each of the three terms in the above expression with

$$\mathcal{G}_n \equiv \Big\{ x\in\mathbb{R}^n \,:\, \max_{i\leq 1}|u_i^*x| \leq \sqrt{6\log n}\Big\} , \qquad (23)$$

Notice that $\mathbb{P}\{(u_i^*x_0)^2 \geq 6\log n\} \leq 2/n^3$. By the union bound we get $\mathbb{P}\{x_0 \in \overline{\mathcal{G}}_n\} \leq 2/n^2$.

Next observe that

$$\frac{1}{\lambda^t}\mathbb{E}\{\widehat{\lambda}^t|\,\mathcal{G}_n\} - 1 = \big(\mathbb{E}\{(u_1^*x_0)^2|\mathcal{G}_n\} - 1\big) + \sum_{i=2}^n \frac{\lambda_i^t}{\lambda^t}(u_i^*x_0)^2 .$$

The first step can be computed as

$$\mathbb{E}\{(u_1^*x_0)^2|\mathcal{G}_n\} = \frac{\mathbb{E}\{(u_1^*x_0)^2\} - \mathbb{E}\{(u_1^*x_0)^2\,\mathbb{I}_{\overline{\mathcal{G}}_n}\}}{1 - \mathbb{P}\{x_0\in\overline{\mathcal{G}}_n\}} ,$$

whence, recalling that $\mathbb{E}\{(u_1^*x_0)^2\} = 1$, and $\mathbb{P}\{x_0 \in \overline{\mathcal{G}}_n\} \leq 1/2$ for all n large enough, we get

$$\begin{aligned} \big|\mathbb{E}\{(u_1^*x_0)^2|\mathcal{G}_n\} - 1\big| &= \left|\frac{\mathbb{P}\{x_0\in\overline{\mathcal{G}}_n\} - \mathbb{E}\{(u_1^*x_0)^2\,\mathbb{I}_{\{x_0\in\overline{\mathcal{G}}_n\}}\}}{1-\mathbb{P}\{x_0\in\overline{\mathcal{G}}_n\}}\right| \\ &\leq \frac{4}{n^2} . \end{aligned}$$

Note further that, by Chernoff inequality, $\sum_{i=1}^n (u_i^* x_0)^2 \leq 3n$ with probability at least $1 - \exp\{(1/10)n\}$. Then,

$$\Big| \frac{1}{\lambda^t} \mathbb{E}\{\widehat{\lambda}^t \big| \mathcal{G}_n\} - 1 \Big| \leq \frac{4}{n^2} + 3n \Big(\frac{|\lambda_2|}{\lambda} \Big)^t .$$

Finally, using Lemma 7.2 and 7.3 we get

$$\frac{1}{\lambda^{2t}} \operatorname{Var}(\widehat{\lambda}^t | x_0) \leq 4\, n^2 \Big\{ \frac{\alpha^t}{4n} + \frac{2\alpha^2 \gamma}{n^2} + 2 \frac{\alpha \sqrt{\gamma}}{n} \sum_{i=1}^n \frac{|\lambda_i|}{\lambda} (u_i^* x_0)^2 + \frac{\alpha t \gamma}{n} \Big(\sum_{i=1}^n \frac{|\lambda_i|}{\lambda} (u_i^* x_0)^2 \Big)^2 \Big\} .$$

Further, for any $x_0 \in \mathcal{G}_n$, $\sum_{i=1}^n (|\lambda_i|/\lambda)(u_i^* x_0)^2 \leq 6\, \gamma \log n$, and therefore

$$\begin{aligned}
&\frac{1}{\lambda^{2t}} \operatorname{Var}(\widehat{\lambda}^t | x_0) \\
&\leq 4n \Big\{ \frac{\alpha^t}{4} + \frac{2\alpha^2 \gamma}{n} + \frac{12 \alpha \gamma^{3/2} \log n}{n} + \frac{36 \alpha t \gamma^3 (\log n)^2}{n^2} \Big\} \\
&\leq 4n \Big\{ \frac{3\alpha^2 \gamma}{n} + \frac{12 \alpha \gamma^{3/2} (\log n)}{n} + \frac{36 \alpha t \gamma^3 (\log n)^2}{n^2} \Big\} \\
&\leq 4 \Big\{ 15 \alpha \gamma^{3/2} (\log n) + \frac{36 \alpha t \gamma^3 (\log n)^2}{n} \Big\} ,
\end{aligned}$$

where we used the fact that, for $\alpha < 1/2$ and $t \geq \log_2(n)$, we have $\alpha^t/4 \leq \alpha^2 \gamma / n$.

Collecting the various terms we obtain

$$\begin{aligned}
&\mathbb{P}\big\{ \widehat{\lambda} \notin [\lambda(1-\Delta)^{1/t}, \lambda(1+\Delta)^{1/t}] \big\} \\
&\leq \frac{2}{n^2} + \frac{1}{\Delta^2} \Big\{ \frac{4}{n^2} + n l_2^t \Big\}^2 + \frac{4\alpha \log n}{\Delta^2} \Big\{ 15 \gamma^{3/2} + \frac{36 t \gamma^3 (\log n)}{n} \Big\} ,
\end{aligned}$$

whence $\mathbb{P}\big\{ \widehat{\lambda} \notin [\lambda(1-\Delta)^{1/t}, \lambda(1+\Delta)^{1/t}] \big\} \leq \delta$, provided $n \geq 4/\sqrt{\delta}$, $\Delta \geq 4\sqrt{2}/(n\sqrt{\delta})$, $\Delta \geq 2n(|\lambda_2|/\lambda)^t/\sqrt{\delta}$, $\Delta^2 \geq (240 \alpha \gamma^{3/2} \log n)/\delta$, $\Delta^2 \geq 4 \cdot 144 \alpha t \gamma^3 (\log n)^2/(n\delta)$. The thesis follows by noting that $(1+\Delta)^{1/t} \leq 1 + (\Delta/t)$ and $(1 - \Delta)^{1/t} \geq 1 - 2(\Delta/t)$ provided $\Delta \leq 1/2$.

8. REFERENCES

[1] Ekahau. http://www.ekahau.com.

[2] Qwikker. http://qwikker.com.

[3] Sonitor technologies. http://www.sonitor.com.

[4] D. Achlioptas and F. McSherry. Fast computation of low-rank matrix approximations. *J. ACM*, 54(2):9, 2007.

[5] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Trans. on Inform. Theory*, 52:2508 – 2530, 2006.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30:107–117, April 1998.

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for information science*, 41(6):391–407, 1990.

[8] A. Dimakis, S. Kar, J. Moura, M. Rabbat, and A. Scaglione. Gossip algorithms for distributed signal processing. *Proc. of the IEEE*, 98:1847–1864, 2010.

[9] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering in large graphs and matrices. In *SODA '99*, pages 291–299, 1999.

[10] P. Drineas and R. Kannan. Pass efficient algorithms for approximating large matrices. In *SODA '03*, pages 223–232, 2003.

[11] P. Drineas, R. Kannan, and M. W. Mahoney. Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix. *SIAM J. Comput.*, 36(1), 2006.

[12] P. Drineas and M. W. Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.*, 6:2153–2175, 2005.

[13] P. Frasca, R. Carli, F. Fagnani, and S. Zampieri. Average consensus by gossip algorithms with quantized communication. In *47th IEEE Conference on Decision and Control*, pages 4831–4836, 2008.

[14] A. Frieze, R. Kannan, and S. Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *J. ACM*, 51(6):1025–1041, 2004.

[15] H. Furstenberg and H. Kesten. Products of random matrices. *The Annals of Mathematical Statistics*, 31(2):457–469, June 1960.

[16] N. Halko, P. Martinsson, and J. A. Tropp. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. `arXiv:0909.4061`, 2010.

[17] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, 2003.

[18] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[19] I. T. Jolliffe. *Principal component analysis*. Springer-Verlag, 1986.

[20] A. Kashyap, T. Basara, and R. Srikant. Quantized consensus. *Automatica*, 43:1192–1203, 2007.

[21] D. Kempe and F. McSherry. A decentralized algorithm for spectral analysis. *Journal of Computer and System Sciences*, 74(1):70 – 83, 2008.

[22] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Trans. Inform. Theory*, 56(6):2980–2998, June 2010.

[23] M. Le Page. Theoremes limites pour les produits de matrices aleatoires. *Probability Measures on Groups*, 928:258–303, 1982.

[24] S. Oh, A. Karbasi, and A. Montanari. Sensor network localization from local connectivity: Performance analysis for the mds-map algorithm. In *Proc. of the IEEE Inform. Theory Workshop*, January 2010.

[25] V. Oseledets. A multiplicative ergodic theorem. lyapunov characteristic numbers for dynamical systems. *Trans.Moscow Math. Soc.*, 19:197–231, 1968.

[26] B. W. Parkinson and J. J. Spilker. *The global positioning system: theory and applications*. American Institute of Aeronautics and Astronautics, 1996.

[27] D. Shah. Gossip algorithms. *Foundations and Trends in Networking*, 3, 2009.

[28] D. Spielman and N. Srivastava. Graph sparsification by effective resistances. In *40th annual ACM symposium on Theory of computing*, 2008.

[29] J. B. Tenenbaum, V. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

APPENDIX

A. PROOF OF THEOREM 5.1

We start by restating the main result of [23], in a somewhat more explicit form. Recall that $f : \mathsf{P}_n \to \mathbb{R}$ is said to be λ-Hölder continuous if its *Hölder coefficient*, defined by

$$[f]_\lambda = \sup_{\mathbf{x}\neq\mathbf{y}} \frac{|f(\mathbf{x}) - f(\mathbf{y})|}{d(\mathbf{x},\mathbf{y})^\lambda}\,, \tag{24}$$

is finite.

THEOREM A.1 (LE PAGE, 1982). *Under assumptions* L1 *and* L2 *there exists a unique measure μ on P_n that is stationary for the Markov chain $\{\mathbf{X}_t\}$. Further, there exists constants $A \geq 0$, $\rho \in (0,1)$, $\lambda \in (0,1]$ such that, for any λ-Hölder function $f : \mathsf{P}_n \to \mathbb{R}$,*

$$\big|\mathbb{E}\{f(\mathbf{X}_t)\} - \mu(f)\big| \leq A\,\rho^t\,[f]_\lambda\,.$$

Remark: The above follows immediately from Theorem 1 in [23] via a simple coupling argument. Notice in particular that it applies to any Lipschitz function since $[f]_\lambda$ is upper bounded by the Lipschitz modulus of f.

Next we restate and prove the first part of Theorem 5.1.

THEOREM A.2. *Assume conditions* L1 *and* L2 *hold, together with* A1, A2. *Denote by μ the unique stationary measure of the Markov chain $\{\mathbf{X}_t\}_{t\geq 0}$. Then*

$$\mu(\mathsf{G}^c) = 0. \tag{25}$$

PROOF. Consider a Markov chain MC_1 with $\mathbf{x}_0 \in \mathsf{G}$. The Markov chain MC_1 has a stationary distribution because conditions L1 and L2 hold. From the property A1, we know that $\mathbf{x}_t \in \mathsf{G}$. Therefore the stationary distribution of MC_1, say μ_1, satisfies $\mu_1(\mathsf{G}^c) = 0$. From Theorem A.1 we know that the stationary distribution is unique. Therefore $\mu = \mu_1$ and hence $\mu(\mathsf{G}^c) = 0$. □

Finally, we will state and prove a generalization of the second part of Theorem 5.1. For this we generalize hypothesis A2 as follows.

A2'. For any $\mathbf{x} \neq \mathbf{y} \in \mathsf{G}$, $\mathbb{E}\left[d(S^{(t)}\mathbf{x}, S^{(t)}\mathbf{y})^\lambda\right] \leq \rho\, d(\mathbf{x},\mathbf{y})^\lambda$.

THEOREM A.3. *Assume conditions* L1 *and* L2 *hold, together with* A1 *and* A2'. *Denote by μ the unique stationary measure of the Markov chain $\{\mathbf{X}_t\}_{t\geq 0}$. Let $\mathbf{x}_0 \in \mathsf{G}$. Then for any λ-Hölder function $f : \mathsf{P}_n \to \mathbb{R}$, we have*

$$\big|\mathbb{E}\{f(\mathbf{X}_t)\} - \mu(f)\big| \leq \rho^t\,[f]_\lambda\,. \tag{26}$$

The proof of this theorem is based on a coupling argument. The coupling assumed throughout is fairly simple: given initial conditions $\mathbf{x}_0, \mathbf{y}_0 \in \mathsf{G}$, we define the chain $\{(\mathbf{X}_t, \mathbf{Y}_t)\}_{t\geq 0}$ by letting $(\mathbf{X}_0, \mathbf{Y}_0) = (\mathbf{x}_0, \mathbf{y}_0)$ and, for all $t \geq 1$,

$$\mathbf{X}_t = S^{(t)}S^{(t-1)}\cdots S^{(1)}\mathbf{x}_0\,, \qquad \mathbf{Y}_t = S^{(t)}S^{(t-1)}\cdots S^{(1)}\mathbf{y}_0\,. \tag{27}$$

It is further convenient to introduce, for $t \in \mathbb{N}$, $\lambda > 0$ the quantity

$$\rho_\lambda(t) \equiv \sup_{\mathbf{x}_0\neq\mathbf{y}_0\in\mathsf{G}} \mathbb{E}\Big\{\Big[\frac{d(\mathbf{X}_t,\mathbf{Y}_t)}{d(\mathbf{X}_0,\mathbf{Y}_0)}\Big]^\lambda\Big\}\,. \tag{28}$$

PROOF. First notice that the function $t \mapsto \rho_\lambda(t)$ is sub-multiplicative. This follows from

$$\begin{aligned}
\rho_\lambda(t_1+t_2) &= \sup_{\mathbf{x}_0\neq\mathbf{y}_0\in\mathsf{G}} \mathbb{E}\Big\{\Big[\frac{d(\mathbf{X}_{t_1+t_2},\mathbf{Y}_{t_1+t_2})}{d(\mathbf{X}_0,\mathbf{Y}_0)}\Big]^\lambda\Big\} \\
&= \sup_{\mathbf{x}_0\neq\mathbf{y}_0\in\mathsf{G}} \mathbb{E}\Big\{\Big[\frac{d(\mathbf{X}_{t_1},\mathbf{Y}_{t_1})}{d(\mathbf{X}_0,\mathbf{Y}_0)}\Big]^\lambda\Big[\frac{d(\mathbf{X}_{t_1+t_2},\mathbf{Y}_{t_1+t_2})}{d(\mathbf{X}_{t_1},\mathbf{Y}_{t_1})}\Big]^\lambda\Big\} \\
&\overset{(a)}{\leq} \sup_{\mathbf{x}_0\neq\mathbf{y}_0\in\mathsf{G}} \mathbb{E}\Big\{\Big[\frac{d(\mathbf{X}_{t_1},\mathbf{Y}_{t_1})}{d(\mathbf{X}_0,\mathbf{Y}_0)}\Big]^\lambda\Big\} \sup_{\mathbf{x}_0\neq\mathbf{y}_0\in\mathsf{G}} \mathbb{E}\Big\{\Big[\frac{d(\mathbf{X}_{t_2},\mathbf{Y}_{t_2})}{d(\mathbf{X}_0,\mathbf{Y}_0)}\Big]^\lambda\Big\} \\
&= \rho_\lambda(t_1)\rho_\lambda(t_2)\,.
\end{aligned}$$

where (a) follows from the condition A1. From the condition A2', we know that $\rho_\lambda(1) \leq \rho$, hence $\rho_\lambda(t) \leq \rho^t$.

Next let $\{\mathbf{X}_t\}_{t\geq 0}$ and $\{\mathbf{Y}_t\}_{t\geq 0}$ be Markov chains coupled as above, with initial conditions $\mathbf{X}_0 = \mathbf{x}_0 \in \mathsf{G}$ and $\mathbf{Y}_0 \sim \mu$. We then have

$$\begin{aligned}
\big|\mathbb{E}\{f(\mathbf{X}_t)\} - \mu(f)\big| &= \big|\mathbb{E}\{f(\mathbf{X}_t)\} - \mathbb{E}\{f(\mathbf{Y}_t)\}\big| \\
&\leq \mathbb{E}\big\{|f(\mathbf{X}_t) - f(\mathbf{Y}_t)|\big\} \\
&\leq [f]_\lambda \mathbb{E}\big\{d(\mathbf{X}_t,\mathbf{Y}_t)^\lambda\big\} \\
&\leq [f]_\lambda \mathbb{E}\Big\{\Big[\frac{d(\mathbf{X}_t,\mathbf{Y}_t)}{d(\mathbf{X}_0,\mathbf{Y}_0)}\Big]^\lambda\Big\} \\
&\leq [f]_\lambda\,\rho_\lambda(t) \leq [f]_\lambda \rho^t\,,
\end{aligned}$$

where we used $d(\mathbf{X}_0,\mathbf{Y}_0) \leq 1$ by definition. This concludes our proof. □