Highly Scalable Linear Time Estimation of Spectrograms - A Tool for Very Large Scale Data Analysis

O. Bhardwaj
Rensselaer Polytechnic Institute
12180 Troy, USA
bhardo@rpi.edu

Y. Ineichen, C. Bekas, and A. Curioni
IBM Research – Zurich
Computational Sciences
Säumerstrasse 4, 8803 Rüschlikon - Switzerland
{yin,bek,cur}@zurich.ibm.com

When humans are faced with the task to visually interpret and compare data, for example the two leftmost graphs in Fig. 1, we are at a loss. In fact, both graphs are exactly the same as one of the graphs can be transformed into the other by a set of simple permutations.

Figure 1: It is almost impossible to determine if the graphs (left and middle column) are the same for a human observer.

Graphs hold a central role in Big Data analytics. Indeed, graphs and networks are key in the new era of on-line social life, but in addition they are increasingly becoming powerful tools in diverse fields such as genomics, systems biology and financial risk analysis. Analysis of graphs becomes quite a formidable task simply because of their shear size in important applications.

Here we propose a new tool: graph spectrograms, that is the density of eigenvalues of the adjacency matrix of the graph in a certain unit of space. Indeed, the spectrogram shown on the right in Fig. 1 captures the essential characteristic of the graphs immediately and graphically. It transforms complex graphs into a 1-dimensional vector - a simple picture that fosters convenient interpretation.

How about comparing graphs that are almost similar (see Fig. 2)? In essence, the spectrogram provides us with a human readable (low dimensional) clue. Of course, this is a massive dimensionality reduction, however the shape of the spectrogram yields a tremendous wealth of information.

Solving the resulting eigenvalue problems is getting much harder to master in the era of big data. The cubic complexity of dense methods and the limitation of iterative techniques to look deep into the interior of the spectrum at an acceptable cost, call for a new approach.

Method. In the following we present our method to estimate the spectrogram of any real symmetric sparse matrix $A \in \mathbb{R}^{n \times n}$. We start by estimating $\lambda_{\min}$ and $\lambda_{\max}$ of $A$ by a few steps of Lanczos, in order to shift and scale the matrix $A$ to have its spectrum in the interval $[-1, 1]$. Next, we divide the range $[-1, 1]$ in $b$ bins $\mu$, known as inflection points. Subsequently we have to estimate the number of eigenvalues below $\mu$ to compute the spectrogram.

Consider the eigen-expansion of matrix $A$:

$$ A = Q \Lambda Q^T, \quad (1) $$

where $Q$ is the orthonormal matrix of eigenvectors of $A$ and $\Lambda$ is a diagonal matrix with the eigenvalues of $A$ as its diagonal entries. Now, the trace of the orthogonal projector $Q_{\mu} Q_{\mu}^T$ to the eigenspace of eigenvalues $\lambda_i \leq \mu$ is equal to the trace of the Fermi-Dirac distribution function,

$$ f_{\mu}(\epsilon) = \lim_{T \to 0} \left[ 1 + \exp \left( \frac{\epsilon - \mu}{k_B T} \right) \right]^{-1}. \quad (2) $$

When $T \to 0$ the Fermi-Dirac distribution function approaches the Heaviside step function which takes value 1 for $\epsilon \leq \mu$ and 0 for $\epsilon > \mu$. We will slightly abuse the notation and denote the distribution in the limit as $f_{\mu}(\cdot)$. When applied to a matrix $A$, the Fermi-Dirac distribution is equivalent to computing

$$ f_{\mu}(A) = \left( 1 + \exp \left( \frac{A - \mu I}{k_B T} \right) \right)^{-1}. \quad (3) $$

By applying the Fermi-Dirac distribution function $f_{\mu}(\cdot)$ to
both sides of equation (1), we obtain
\[ f_s(A) = f_s(QAQ^T) = Q f_s(A) Q^T. \] (4)
As \( T \to 0 \), the matrix \( f_s(A) \) approaches a diagonal matrix with zeros on the diagonal, except for the entries where \( \lambda_{ij} < \mu \), where it approaches 1. In combination with the orthonormality of \( Q \) we obtain
\[ tr(f_s(A)) = \sum_{i=0}^{n} I_{\{\lambda_i < \mu\}} \] (5)
from equation (4). The last equation tells us how to estimate the number of eigenvalues of \( A \) in the interval \([a, b]\): we just need to estimate the diagonals of \( f_s(A) \) and \( f_s(A) \). The result follows a simple subtraction. The presented mathematical framework can be extended to estimate the spectrogram of \( A \) partitioned into \( b \) bins. Therefore, the problem of estimating the spectrogram of \( A \) partitioned into \( b \) bins can be reduced to estimating \( tr(f_s(A)) \) where \( \mu \in \{p_1, \ldots, p_b\} \) and \( p_i \) denotes the endpoint of the \( i \)-th bin. Recall that \( f_s(A) \) denotes Fermi-Dirac in the limit (as \( T \to 0 \)), and has the form of the Heaviside step function. It is alluring to approximate the Heaviside step function by a Chebyshev expansion, i.e., Chebyshev polynomials with Jackson smoothing \([4]\)
\[ tr \left( \frac{\alpha_0}{2} I + \sum_{m=0}^{M} g_{m,M} \alpha_m T_m(A) \right) \approx \sum_{i=0}^{n} I_{\{\lambda_i < \mu\}}, \] (6)
where \( \{T_m(\cdot)\}_{m=0}^{M} \) are the first \( M \) Chebyshev polynomials of the first kind.

In combination with the stochastic diagonal estimator described in [2], the trace of (6) can be estimated,
\[ D(F(A)) = \left[ \sum_{k=0}^{n} v_k \odot F(A) v_k \right] \otimes \left[ \sum_{k=0}^{n} v_k \cdot v_k \right], \] (7)
where \( \odot \) denotes element-wise dot product, \( \otimes \) denotes element-wise division and \( \{v_k\}_{k=0}^{n} \) are \( n \) i.i.d. random vectors in \( \{-1, 1\}^n \). This yields
\[ F(A) = \frac{\alpha_0}{2} I + \sum_{m=0}^{M} g_{m,M} \alpha_m T_m(A). \] (8)
Using (6), (7), and (8), we see that \( \sum_{i=0}^{n} I_{\{\lambda_i < \mu\}} \) can be computed using \( \mathcal{O}(M) \) matrix-vector products. Therefore, the complexity for computing the \( b \) bin spectrogram is \( \mathcal{O}(b \cdot n^2) \) for a sparse and \( \mathcal{O}(b \cdot n^3) \) for dense real symmetric matrix.

The method described above provides parallelization opportunities on three hierarchical levels. The method relies on: (i) independent computation of the bins, (ii) independent matrix-vector multiplications for trace estimation, and (iii) scalable matrix-vector multiplications.

Results. For experimental purposes, we converted matrices provided by the UFLORIDA\(^1\) sparse matrix collection to unweighted graphs, ignoring self-loops. To validate our spectrogram approach we chose the “laser” matrix, a small matrix of size 3002 x 3002 with 9000 non-zeros, where all eigenvalues can be computed explicitly (NumPy). As shown
\[\text{Fig. 3: Validation for laser.mtx using 50 bins with } s = 256. \text{ Left: } M = 32, \text{ right: various } M.\]

We evaluated the outlined method on a medium graph, namely the bmwca_1 matrix (dimensions 148’770 x 148’770 and \( 10’641’602 \) non-zeros) for 8192 bins with \( M = 64 \) and \( s = 64 \). The scalability is evaluated on a BlueGene/Q system\(^2\). Fig. 4 shows the excellent scalability of the spectrogram estimation technique.

\[\text{Fig. 4: Speedup for average of total runtime with 8192 bins, } M = 64, \text{ and } s = 64 \text{ on a BG/Q.}\]

Conclusions. Graph spectrograms hold a wealth of information and can open new ways for analytics. Our work renders spectrograms to be a powerful tool for mining Big Data, by tackling large scale problems. Graphs in the order of millions require just a few seconds on modest computational resources, while standard methods would already need Exascale machines. We have just begun to scratch the surface. We place a bet for SC14 in NOLA: To show the world spectrograms of Internet scale graphs.

1 REFERENCES

\(^{1}\)IBM, ibm.com, and Blue Gene/Q are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide.