Distributed Spectral Dimensionality Reduction for Visualizing Textual Data

Sanjay Krishnan SANJAY@EECS.BERKELEY.EDU
Electrical Engineering and Computer Sciences
University of California-Berkeley

Ken Goldberg GOLDBERG@BERKELEY.EDU
IEOR, EECS, and School of Information
University of California-Berkeley

Abstract

We use a Spectral Clustering model to formulate a distributed implementation using SPARK of Laplacian Eigenmaps that we call Distributed Spectral Dimensionality Reduction (DSDR). We evaluate DSDR to visualize conceptual clusters of terms in textual data from 2149 short documents written by online contributors to a State Department website. We compare DSDR with PCA, Multi-Dimensional Scaling, ISOMAP, and Locally Linear Embedding based on the Dunn Separation Index and computation times. We find for this dataset that DSDR is faster and better preserves high-dimensional cluster structure.

1. Introduction

Spectral clustering methods have shown empirical success in image segmentation, on example datasets with non-convexities, and document grouping for information retrieval (Shi and Malik, 2000; Ng et al., 2001; He et al., 2011; Srivatsa, 2005). Recent work on the subject has explored scalability solutions such as approximations, an approximate HADOOP implementation, and a parallellized implementation (Chen and Cai, 2011; Yan et al., 2009; Hefeeda et al., 2012; Chen et al., 2011). We build on those insights and designed a scalable spectral dimensionality reduction technique for large-scale data exploration and visualization. This work is motivated by a relaxation of the NP-Complete

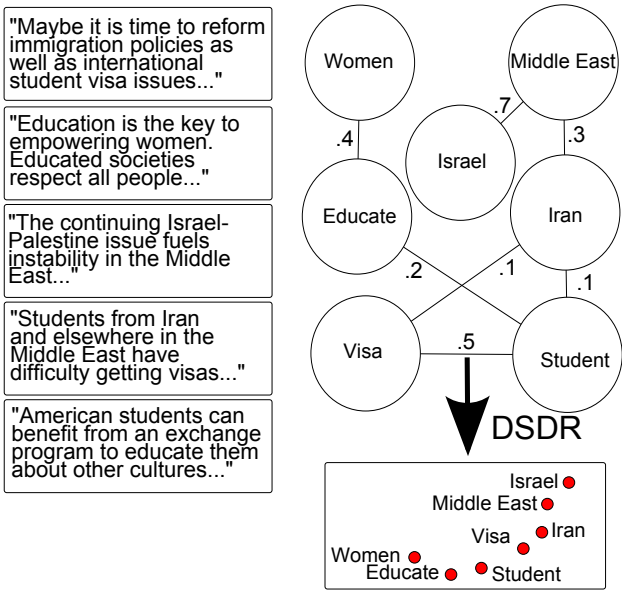

Figure 1. Schematic representation of the model and algorithm. DSDR efficiently takes a graphical representation of correlation relations between terms, and forms a representative 2D map.

normalized min-cut graph partitioning problem, variants of which appear in VLSI design, bi-partite matching, and computational geometry (Alpert et al., 1995; Dhillon, 2001; Spielman et al., 1996). Belkin and Niyogi showed that this relaxation, involving a generalized eigenvector problem of the graph Laplacian matrix, constituted a well-defined dimensionality reduction which they termed, Laplacian Eigenmaps (Belkin and Niyogi, 2003). We propose a new algorithm, Distributed Spectral Dimensionality Reduction (DSDR), derived from Laplacian Eigenmaps but optimized for a distributed implementation.

We implement this algorithm on SPARK, a

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Submitted to ICML Spectral Workshop 2013

MapReduce-like distributed platform that uses an intelligent task management system and in-memory computation to address the inefficiencies observed in prior MapReduce implementations such as HADOOP (Bu et al., 2010). Increasing memory and coordinated task scheduling is a new trend in computing clusters, so emerging approaches, such as SPARK, revise the traditional MapReduce computational model to one that is appropriate for these systems and can support complex Machine Learning tasks (Zaharia et al., 2011).

To evaluate DSDR, we experiment with textual data from the collaborative brainstorming platform, Opinion Space (Faridani et al., 2010). In 2010, a version of Opinion Space was used by the U.S. State Department, where it attracted thousands of participants to contribute suggestions on foreign policy. Particpants responded to the discussion question: *If you met U.S. Secretary of State Hillary Clinton, what issue would you tell her about, why is it important to you, and what specific suggestions do you have for addressing it?* We work with a datset of 2149 short textual documents contributed by the participants and apply DSDR to find 2D visualization terms that extracts the latent conceptual groupings of terms (topics).

Dimensionality reduction has long been a key component of topic modeling algorithms. In the seminal work on Latent Semantic Analysis (Deerwester et al., 1990) (also known as LSI in Information Retreival literature), the problem was formulated as a low-rank approximation of the term-document matrix. For N documents and a dictionary of D terms, they define an N by D term-document matrix A that indicates the presence of a term in a given document with a non-zero value. The top k eigenvectors of A^TA form a k-dimensional subspace of terms, and they interpret each of the vectors as a latent topic or concept. The process of taking the top k eigenvectors is equivalent to a k-dimensional Principal Component Analysis (PCA) projection of terms. At its core, LSA is using a dimensionality reduction to infer a high dimensional cluster structure of words. With this in mind, our experiments suggest that DSDR preserves high dimensional clustering of the textual data better than PCA and improved topic extraction compared to LSA.

2. Related Work

The representative approaches proposed in recent work (Yan et al., 2009; Chen and Cai, 2011) point to the main bottleneck in spectral clustering, forming the similarity graph. In Distributed Approximate Spectral Clustering (DASC) (Hefeeda et al., 2012), the authors argue for a Locality Sensitive Hashing heuristic for neighbor finding and a HADOOP solution to the formation problem. DASC uses a QR+Lanczos iteration approach to spectral decomposition, while we use a different formulation of the problem and apply a simpler power-iteration workflow that can take advantage of SPARK's computational model.

In user interface design, others methods have been applied for term-space visualization. The WordSpace project (Brandes et al., 2006) visualizes the space of words with force-directed layouts, a graph visualization algorithm that treats weighted edges like a forces on vertices and finds a layout that is in equilibrium. (Etling et al., 2009) developed a linguistic map of the Middle Eastern blogosphere, based on the Fruchterman-Rheingold layout algorithm of a link graph. These projects did not emphasize mathematics of the graph embedding or topic modeling, and the techniques selected are actually equivalent to metric Multi-Dimensional Scaling.

3. Similarity Model for Text Analysis

In LSA term-document model A^TA is a D by D matrix whose ij^{th} element is an inner product; a measure of the covariance of two terms i and j. This similarity model is biased towards frequent terms so often a Term Frequency Inverse Doucment Frequency(tf-idf) transformation is applied factor out this effect. To apply DSDR, we interpret a mean-centered A^TA as representing a complete graph of D terms assigning each edge a weight of $cov(i, j)$. We sparsify the graph by removing the edges when the correlation is negative and or statistically insignificant based on a Student's t null-hypothesis test $p > .05$.

Due to its normalized min-cut derivation and the sparse model, DSDR can use the same comparision model without the need of tf-idf. In our formulation of Laplacian Eigenmaps, the edges are normalized by the cummulative edge weights at source, and the sparsity constraints retain only the statistically significant edges. By normalizing the edges and reducing the degree of the vertices corresponding to the most frequent terms, their spectral contribution is reduced. In many neighbor-based approaches, determining the size of the neighborhood is an important design parameter and our statistical significance method automatically chooses a sparse model. In practice, this test gives us a threshold based on the number of documents N in the corpus of $\epsilon \approx \frac{2}{\sqrt{N}}$ and can be interpreted as connecting to all terms within a $1 - \epsilon$ radius spherical neighborhood.

Algorithm 1 DSDR

1. Partition data into equal-sized blocks.
2. MAP over each pair of blocks, calculate the covariance matrix.
3. FILTER edge weights that pass a statistical significance test $p < .05$, are positive, or on the diagonal.
4. REDUCE over all vertices to find the cumulative weight of edges connected to each vertex.
5. MAP over all edges normalizing thet weigh by the cumulative weight at each endpoint, and let W be the adjacency matrix of this graph.
6. Calculate the right eigenvectors of W with power-iteration. Starting from the second largest eigenvector take the next k.

4. Algorithm and Distributed Implementation

Belkin and Niyogi's Laplacian Eigenmaps calculates the k-dimensional embedding by solving the generalized eigenvalue problem $Lf = \lambda Vf$ for the k smallest non-trivial eigenvectors, where L is the graph Laplacian matrix and V is a diagonal matrix where V_{ii} is the cumulative flow out of a vertex i. We reformulate the generalized eigenvector problem to dominant eigenvector problem for $V^{-1}W$ with W a weight matrix where W_{ij} is the weight of the edge between vertex i and j, and apply a MAP-REDUCE-BROADCAST power-iteration to solve for the eigenvectors. Power-iteration converges quickly for a sparse matrix and for N documents and D terms, the time complexity of our algorithm is $O((N + k)D^2)$ with $N >> k$. Power-iteration also allows us to incrementally calculate the eigenvectors as neeeded, and in an application such as visualization this means only the 3 dominant eigenvectors (the dominant eigenvector corresponds to a trivial cut).

The time complexity is clearly dominated by the graph formation step and this is easliy distributable. In contrast to prior approaches in distributed spectral clustering, we work on blocks of data rather than individual data vectors; distributing blocks in pairs to the nodes in our cluster. We can use efficient linear algebra primitives incurring communication overheads once instead of element-wise operations.

5. Results

The dataset we used consisted of 2,149 textual documents each with at most 1,000 characters, and a total 11,817 unique terms that appeared in at least 5 documents or more. We manually cleaned the dataset for common entity resolution problems eg. resolving "USA", "U.S.", "America", and "U.S.A" together, and removed stop-words based on the list at (http://jmlr.csail.mit.edu). Empirically, we found that the similarity graphs was sparse with the average vertex connected to only 26.86 ± 9.09 other vertices, and every vertex was connected to at least one other vertex. In Figure 2(included at the end), we visualize the 200 most frequent nouns as an example.

5.1. Method Comparison

We evaluate our method (single-node implementation) against other dimensionality reduction schemes on how well the techniques preserve clustering properties after dimensionality reduction. We quantify the quality of clustering with the Dunn Separation Index (Dunn, 1974), which is defined as the ratio of the minimum distance between a point in the cluster and one outside and maximum distance between points within a cluster. The intuition behind this index is that clusterings that assign dense clusters that are far apart lead to higher values, and the index is independent of metric space distortions that happen with dimensionality reduction. Since we apply DSDR to visualization, we experiment with a method comparision of 2D dimensionality reductions.

We evaluated the Dunn Index of a k-means clustering (for k = 10) of the 500 terms sampled from high-dimensional data (0.7372) as a baseline, where we selected k as the number of clusters that led to the highest Dunn Index over (k = 1 to k = 50). For four other dimensionality reduction schemes Multi-Dimensional Scaling (MDS), Locally Linear Embedding (LLE), Principal Component Analysis (PCA), and ISOMAP we reduced the data to a 2D space and reclustered using k-means, giving us a new clustering. Based on the original high-dimensional locations of the points, we calculate the Dunn Index of this clustering and to account for the randomized effects of k-means initialization, we report the maximum Dunn index for 100 trials. The resulting metric gives us a measure of how well clusters in the embedded space are separated with relation to their original locations. Finally, we also include the single-node runtime for each method. While we compare the methods on single node performance, there are important parallelization considerations. ISOMAP requires a full all-points shortest path calculation over the similarity graph and thus cannot be parallelized easily. LLE requires a complex k-nearest neighbor query which is difficult to implement with in the MapReduce or SPARK model where communication between nodes is limited.

Technique	Dunn Index	Runtime
MDS	0.2481	45.12 secs
PCA	0.3639	7.23 secs
ISOMAP	0.7003	143.94 secs
LLE	0.6711	26.71 secs
DSDR	0.7150	12.48 secs

Table 1. Dunn Separation Index evaluation of dimensionality reduction and subsequent k-means clustering. We found that PCA and MDS give poor clustering results, compared to DSDR, ISOMAP, and LLE. In addition, we found that for a dataset of 500 words DSDR is faster than the other cluster preserving alternatives.

The cluster preservation properties of DSDR are consistent with the argument made by Lee and Verleysen that Laplacian Eigenmaps are topology preserving even though the technique may distort metric spaces (Lee and Verleysen, 2007). LLE and ISOMAP are slower since they attempt to model the data manifold structure accurately. In this application of DSDR, we find topics through a clustering process which makes cluster preservation a more meaningful analysis than metric space preservation.

5.2. LSA and DSDR

So far, we have been comparing DSDR and LSA by looking at PCA's preservation of cluster structure. We can qualify the effects of a poor clustering with examples from our dataset. In LSA, we typically identify topics by using the largest components of the orthogonal topic vectors, for consistency we apply the same to the DSDR axes.

LSA Topic 1 *settlement, cost, focus, troops, address, violence, research, south, climate, reform.*

LSA Topic 2 *student, country, middle east, visa, palestine, government, state, israel, united states, study.*

LSA Topic 3 *world, education, women, countries, stability, rest, children, food, poverty, development.*

DSDR Topic 1 *multiple, country, year, home, years, family, student, entry, visa, study.*

DSDR Topic 2 *population, country, climate, change, planet, energy, production, development, health, women, food, education.*

DSDR Topic 3 *broker, arab, israel, two, state, conflict, palestine, middle east, peace, settlement.*

Not surprisingly, LSA results indicate less homogenous clusters which confirm intuition about the underlying PCA dimensionality reduction.

5.3. Scalability

We applied our algorithm to the 11,400 most frequent terms in the dataset and ran the experiment on 4 m1.large Amazon EC2 instances. In this experiment, the data was partitioned into equal-sized blocks and stored in common storage between the nodes. When restricted to a single MapReduce task on a single core, the algorithm required 678.93 secs to find a 2D embedding. We expanded that to 4 4-core nodes each core running 2 MapReduce tasks and the entire algorithm required 62.12 secs. We found for this dataset the algorithm scales almost linearly, with the only nonlinear overheads in normalization, and power-iteration broadcasts.

6. Conclusion

We proposed the Distributed Spectral Dimensionality Reduction (DSDR) algorithm for large-scale dimensionality reduction. Experiments suggest that DSDR has valuable clustering preserving properties that make it ideal choice for problems like topic modeling. Our experiments and results tie back in to the normalized min-cut problem, the original motivation for spectral clustering. In this problem, we partition a graph into two disjoint sets A and B by finding a cut that minimizes the sum of edges across the cut normalized by the sum of the edges contained in either A or B. As an objective, it incorporates both internal connectivity and external separability which is very similar to the Dunn index which is the ratio of internal similarity to separation. Not suprisingly, our technique, which can be seen as a soft partitioning of this graph, is very effective at preserving this criteria for clusters.

7. Future Work

Recent work in topic modeling has proposed spectral dimensionality reduction formulation for Latent Dirichlet Allocation (Anandkumar et al., 2012). In this paper, we make geometric argument rather than a statistical one, and comparing our results to the Dirichlet process literation LDA/HDP (Blei et al., 2003; Teh et al., 2006) is a subject of future work. We are also particularly interested in exploring online adaptations of the algorithm applied to visualizing streaming data. We are also collaborating with MLBase team (Kraska et al., 2013) at UC Berkeley to build a general spectral dimensionality reduction toolkit in addition to the SPARK version described in this paper.

The code for this project is available at (https://github.com/BerkeleyAutomation/DSDR)

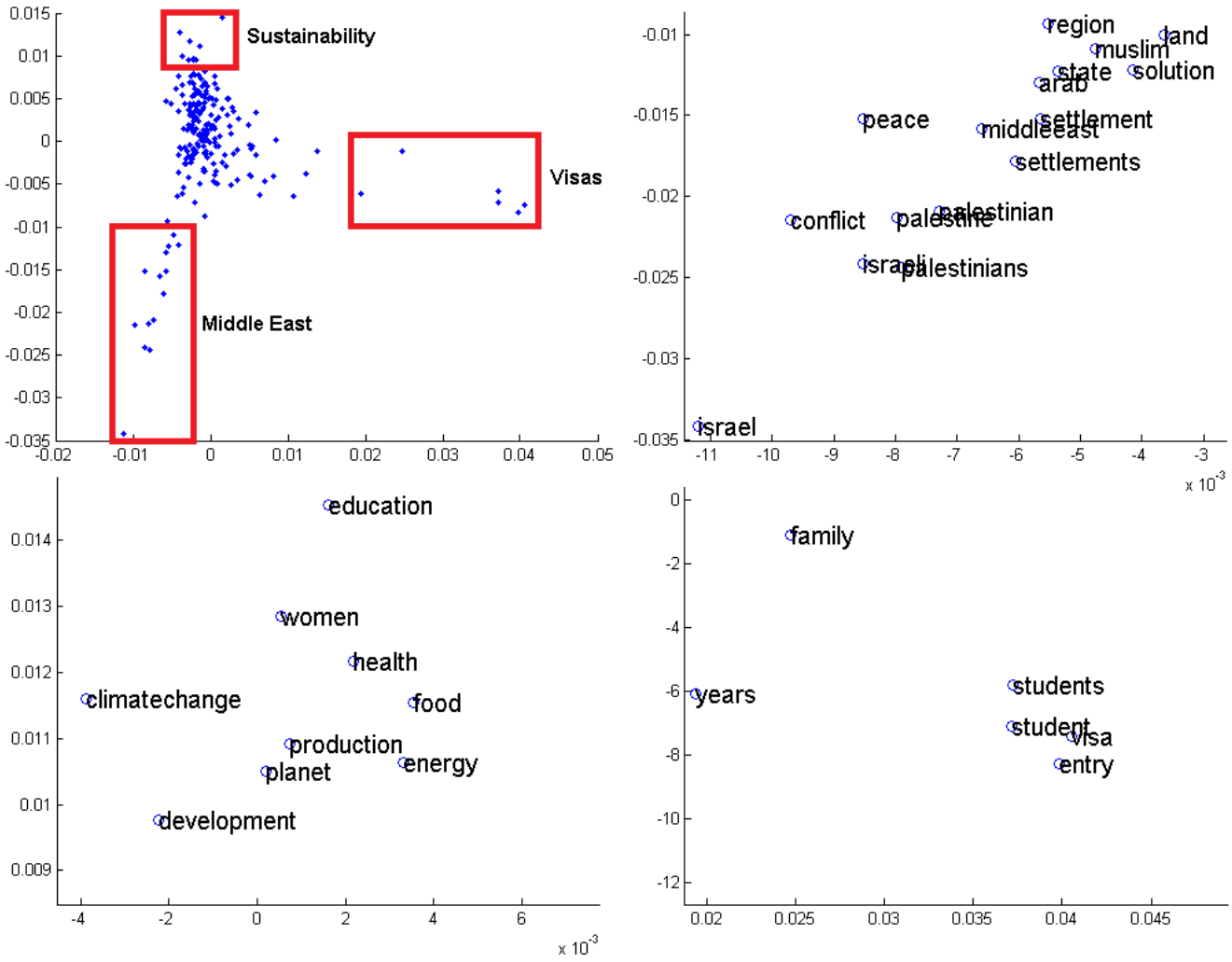

Figure 2. Visualizing the corpus. DSDR finds a 2D embedding from the correlation-based graphical model of the 200 most frequent nouns. The visualization gives us a synopsis of the corpus from which we can find key topics and trends. Our experiments suggest that the embedding preserves high-dimensional clustering, and has improved topic extraction compared to LSA. We highlight the clusters that correspond to the topics described in Section 5.2. *Lower Left*: A region of the 2D plot containing terms relating to student visas and corresponds to DSDR Topic 1. *Lower Right*: A region of the 2D plot containing terms relating to sustainability and climate change and corresponds to DSDR Topic 2. *Upper Right*: A region of the 2D plot containing terms relating to the Middle East and corresponds to Topic 3 in Section 5.2.

References

Alpert, C. and Kahng, A. and Yao, S. *Spectral Partitioning: The More Eigenvectors, The Better* Design Automation, 1995.

Anandkumar, A. and Foster, D. and Hsu, D. and Kakade, S. and Liu, Y. *A spectral algorithm for latent Dirichlet allocation* Advances in Neural Information Processing Systems (NIPS) 25, 2012.

Belkin, M. and Niyogi, P. *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation* Neural Computation, 2003.

Berry, M. and Castellanos, M. *Survey of Text Mining II: Clustering, Classification, and Retrieval* Springer-Verlag, 2008.

Blei, D. and Ng, A. and Jordan, M. *Latent dirichlet allocation* Journal of Machine Learning Research, 2003.

Brandes, U. and Hoefer, M. and Lerner, J. *WordSpace Visual Summary of Text Corpora* 2006.

Bu, Y. and Howe, B. and Balazinska, M. and Ernst, M. *HaLoop: Efficient iterative data processing on large clusters* Proceedings of the VLDB Endowment, 3(1-2), 285-296, 2010.

Chen, X. and Cai, D. *Large Scale Spectral Clustering with Landmark-Based Representation* AAAI, 2011.

Chen, W. *Parallel Spectral Clustering in Distributed Systems* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011.

Deerwester, S. and Dumais, S. and Furnas, G. and Landauer, T.and and Richard, H. *Indexing by latent semantic analysis* JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 1990.

Dhillon, I. *Coclustering documents and words using Bipartite Spectral Graph Partitioning* Knowledge Discovery and Data Mining, 2001.

Dunn, J. *Well separated clusters and optimal fuzzy partitions* Journal of Cybernetics 4: 95104, 1974.

Etling, B. and Kelly, J. and Faris, R. and and Palfrey, J. *Mapping the Arabic Blogosphere: Politics, Culture, and Dissent* Berkman Center Research Publication, 2009.

Faridani, S. and Bitton, E. and Ryokai, K. and Goldberg, K. *Opinion space: a scalable tool for browsing online comments* Proceedings of the 28th international conference on Human factors in computing systems, 2010.

He, X. and Wang, J. and Zhang, Z. and Cai, Y. *Clustering web documents based on Multiclass spectral clustering* International Conference on Machine Learning and Cybernetics (ICMLC), 2011.

Hefeeda, M. and Gao, F. and Abd-Almageed, W. *Distributed approximate spectral clustering for large-scale datasets* 2012.

Huang, L. and Yan, D. and Jordan, M. and Taft, N. *Spectral Clustering with Perturbed Data* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2008.

Kappe, F. and Droschl, G. and Kienreich, and W. and Sabol, V. and Becker, J. and Andrews, K., Granitzer, M., Tochtermann, K., and Auer, P. *InfoSky: Visual Exploration of Large Hierarchical Document Repositories* Proceedings of HCI 2003 International, Creta, Greece, 2003.

Kannan, R. and Vempala, S. and Vetta, A. *On Clusterings - Good, Bad and Spectral* Journal of the ACM, 2000.

Kraska, T. and Talwalkar, A. and Duchi, J. and Griffith, R. and Franklin, M. and Jordan, M. *MLbase:A Distributed Machine Learning System* In Conference on Innovative Data Systems Research, 2013.

Lee, J. and Verleysen, M. *Nonlinear Dimensionality Reduction*. Springer Publishing Company, Incorporated, 1st edition, 2007.

Minier, Z. and Bodo, Z. and Csato, L. *Wikipedia-Based Kernels for Text Categorization* Symbolic and Numeric Algorithms for Scientific Computing, 2007.

Ng, A. and Jordan, M. and Weiss, Y. *On Spectral Clustering: Analysis and an algorithm* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2001.

Papadimitriou, C. and Raghavan, P and Tamaki, H. and Vempala, S. *Latent semantic indexing: A probabilistic analysis* ACM press, 1998.

Paulovich, F. *Text Map Explorer: a Tool to Create and Explore Document Maps* Tenth International Conference on Information Visualization, 2006.

Srivatsa, A. *Discovering recurring anomalies in text reports regarding complex space systems* Aerospace Conference, 2005.

Shi J. and Malik J. *Normalized Cuts and Image Segmentation* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997.

Spielman, D. and Teng, S. *Disk packings and planar separators* In Proceedings of the twelfth annual symposium on Computational geometry (pp. 349-358), 1996.

Teh, Y. and Jordan, M., Beal, J. and Blei, D. *Hierarchical Dirichlet Processes* Journal of the American Statistical Association, 2006.

Yan, D. and Huang, L. and Jordan, M. *Fast Approximate Spectral Clustering* Technical Report No. UCB/EECS-2009-45, 2009.

Zaharia, M. and Chowdhury, M. and Das, T. and Dave, A. and Ma, J. and McCauley, A. and Franklin, M. and Shenker, S. and Stoica, I. *Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing* Electrical Engineering and Computer Sciences University of California at Berkeley Technical Report, 2011.