Modeling Activist Discourse on State Repression Through Text Analytics

Evan Brown and Hailey Reeves
Department of Mathematics, University of Arizona
May 3rd, 2018

Keywords: Sociology, activism, state repression, textual analysis, text mining, latent semantic analysis, network analysis.

ABSTRACT

Our goal is to analyze literature related to state repression from activists within the United States, using a data sample of texts that span across multiple decades and social movements, to uncover trends and perceived costs associated with activist involvement. To analyze our documents, we used text mining - specifically latent semantic analysis and network analysis. By combining these analytical methods, we have revealed links between the documents in our corpus as well as the terms, concepts, and themes contained within them. These underlying patterns may be used by sociologists to create further hypotheses regarding state repression and how it is perceived and handled by American activists.

INTRODUCTION

Our analysis determines common factors associated with activism through studying narratives from activists, specifically focusing on perceived risks associated with engaging in civil disobedience. Through employing multiple textual analysis methods, we measured the most significant influences that discourage activist participation, focussing on deterrents such as state repression, as well as some of the factors that support engagement in the face of repression. We define state repression as the actual or threatened use of penalization against a person or an organization within state jurisdiction, where its purpose is to impose a cost on participants and further prevent activities or beliefs that are perceived to be challenging to state practices or institutions.

We worked directly with Heidi Reynolds-Stenson, a PhD candidate within the Department of Sociology at the University of Arizona. Her dissertation topic is related to how activists perceive and respond to risks of repression. Our data set is a series of texts referred to in transcripts from over fifty Arizona activists she interviewed about direct experiences they have had with repression. The texts are self-published magazines and articles related

to experienced accounts of state repression within social movements, where the authors document examples of events, recovery from, or mitigation of repression.

SUMMARY

After creating frequency histograms and performing latent semantic analysis on our corpus, we constructed both a frequency network and a high degree network, with the latter utilizing the semantic information revealed through LSA. The properties of these are discussed in greater detail in the appropriate section below.

The most important and striking trend revealed in these networks is how both the most frequent and the most connected terms appear to represent layers of opposition present in how activism and state repression are perceived in these documents. The two most frequent terms - people and police - as well as several other terms such as federal and government all represent the two sides of this narrative of opposition, while other frequent and/or central terms represent the methods that either side uses to further their cause - terms such as action, court, information, and law.

LATENT SEMANTIC ANALYSIS

i. Text Mining and Query Matching

In order to create models based on our compendium of texts, we utilized text mining - a blanket term for any process which extracts useful information from a text. Our first step was performing tokenization: the process of converting a stream of text into meaningful elements, called tokens. These tokens could take the form of anything from a single character to an entire phrase, but for our purposes the tokens were singular terms. We created a frequency list for all documents, logging which terms appeared in each document and how often. We then removed unnecessary words, such as conjunctions, articles, and prepositions, known as stop words. Next, we constructed our term-document matrix. A term-document matrix is a matrix where each row corresponds to a term, each column corresponds to a document, and each entry is the frequency with which a term appears in a document.

With the term-document matrix, we were able to perform the process of query matching. First, we introduced a query, a vector corresponding to a set of term frequencies, or to a term itself. To compare this query to other elements in the matrix, we can find the angle between the vectors as a

measure of their similarity. The smaller the angle, the more similar the two vectors are in the frequency and composition of their terms. For a certain document vector aj and a query vector q, the angle between them, θ, can be found with:

$$cos(\theta(q, a_j)) = \frac{q^T a_J}{||q|| * ||a_j||} \tag{1}$$

This formula is easily derived by combining different definitions of the dot product. The algebraic definition for the dot product is:

$$q \cdot a_j = q^T a_j \tag{2}$$

The geometric definition of the dot product is:

$$q^T a_j = ||q|| * ||a_j|| cos(\theta) \tag{3}$$

This can be easily manipulated into the form of (1).

The closer this value of $\cos(\theta)$ is to 1, the smaller the angle, and therefore the more similar the query is to the document vector in question. A similar process can be performed for term vectors. These cosine similarity measures allow us to find patterns in our documents and their terms. Using these measures, we can create a similarity matrix - a grid containing all of our documents used as queries for each other, displaying patterns of similarity

ii. Singular Value Decomposition and k-rank Approximation

While this similarity index does a good job capturing the relationship between documents, we utilized the process of latent semantic analysis to better highlight the similarities between our terms and documents. LSA is an advanced form of query matching that bypasses many problems inherent in basic query matching, extracting the latent semantic information present in a set of documents.

LSA begins with the assumption that words of similar meaning appear in similar contexts, regardless of exact word choice. By comparing these underlying structures instead of exact terms, LSA provides a more accurate measure of the similarity between our documents and terms. To accomplish this, LSA performs a singular value decomposition on the term-document matrix. A singular value decomposition essentially splits a term-document matrix, A_{txd} (where t is the number of terms and d the number of documents, and assuming that t > d), into three different matrices:

$$A_{txd} = T_{txt} S_{txt} (D_{dxt})^T \tag{4}$$

This creates a projection of the information contained within A, where the matrices T and D represent the terms and documents respectively. S is a diagonal matrix containing only the singular values of A, which appear in descending order. From here, we can simplify this matrix by considering only the first k rows, where k is less than t. This is called a k-rank approximation.

There are several different methods for choosing the best value for k, as it is essentially a least-squares approximation problem. The goal is to find the smallest value of k that returns an approximately equal matrix to the original. The method we utilized -using the dimcalc share function in the R package lsa - takes advantage of the fact that singular values are listed in S in descending order. It sums each singular value in order until the sum reaches or exceeds a certain percentage of the total sum of all singular values. Once the threshold is reached, the position of the last used singular value is taken to be the value for k. It is assumed that since the majority of the singular values sum is contained within the approximation, the approximation is sufficiently close to the original to truncate at that dimension. This is because the largest singular values are attached to the most relevant semantic information in our singular value decomposition; focusing on them is what allows to perform more effective similarity tests.

By being able to choose k such that our k-rank approximation is very close to our original matrix, we are left with a matrix that is the same size as the original but is encoded with semantic information about our terms and documents. By performing query matching using this LSA-encoded matrix, a higher cosine value is a better indicator of meaningful semantic similarity than can found by using query matching without LSA.

iii. Results

After processing, our term-document matrix contained 14,642 term entries across each of our 94 documents. LSA was performed on this term-document matrix, with a dimcalc share threshold of 50%. Using this threshold, the optimal k value was found to be 12.

Document similarity matrices were calculated for the term-document matrix both pre-LSA and post-LSA. As the cosine similarity measure always returns a value between 0 and 1, the matrices were color-coded using a red-green gradient, with red indicating less similarity and green indicating more similarity. These are displayed below.

Each matrix has a diagonal line of pure green along its entirety, as the same documents are represented in both the rows and columns, and each document is completely similar to itself. In the pre-LSA matrix (figure 1), one box of green could be seen for documents 72 through 77, showing that these documents all came from the same source, and were focused on the same topic. Similar boxes can be seen for documents 38 through 42 and documents 20 through 30, though only the outlines can be made out, and the similarities are much less distinct. Each group of these documents did, however, come from the same source, so we should see a large amount of similarity between them.

Our post-LSA matrix (figure 2) displays these similarities strikingly, as we can clearly see distinct boxes of similarity in each of the aforementioned sections. After LSA, the documents similarities should be better highlighted, so we should expect to see these source groups more clearly. The presence of these boxes indicates that LSA was successful in highlighting the similarities in our documents.

For individual terms, we were also able to use the associate command to return terms with high similarity values to a query term, above a certain threshold. An example of an associate query can be seen below, in figure 3. With the exception of some erroneous terms, the terms that are returned are semantically similar to the query term. The term rights was closely associated with such terms as allowed, FBI, and organizations, all of which are indeed similar to and associated with the query.

The success of LSA in revealing similarities between both terms and documents indicated that we could use this LSA information to make meaningful networks that were informed by these semantic similarities. This was used to assist in creating networks linking different terms together by their most similar related terms, the details of which will be discussed in the next section.

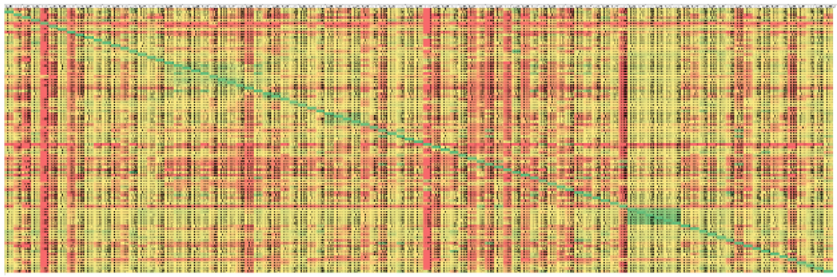

Figure 1: Similarity Matrix, Pre-LSA, with values ranging from 0 to 1 colored on a red-green gradient.

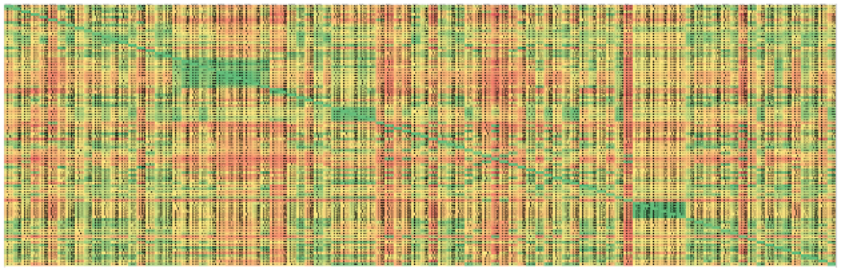

Figure 2: Similarity Matrix, Post-LSA, with values ranging from 0 to 1 colored on a red-green gradient.

```
> associate(lsatdm,"rights",measure = "cosine",threshold=0.85)
government        ing    allowed        fbi
 0.9538582  0.9240945  0.9171399  0.9068765
organizations        usa    lawsuit investigations
 0.8943734  0.8939152  0.8804076  0.8757428
```

Figure 3: Example of associate query on the post-lsa term document matrix, lsatdm. The term rights was queried using a cosine similarity measure, and the command returned all terms with a cosine value greater than 0.85. The terms are listed in descending order, with cosine values below each term.

NETWORK ANALYSIS

From the latent semantic analysis process, high frequency terms generated from the term document matrices were translated into nodes/vertices for the original network. A frequency threshold was defined for the first 20 terms that appeared within a descending frequency. An edge is defined as the terms appearance within a document. Edge weight and direction were not defined. After individual networks were created for each document (for a total of 94 individual graphs), the union was taken for the resulting graph.

Number of vertices:	: 333
Number of edges:	: 656
Average degree:	: 3.9399
Graph density:	: 0.0118
Average clustering:	: 0.0862
Average shortest path:	: 3.8512
Assortativity:	: -0.0481

Figure 4: Frequency Network Properties

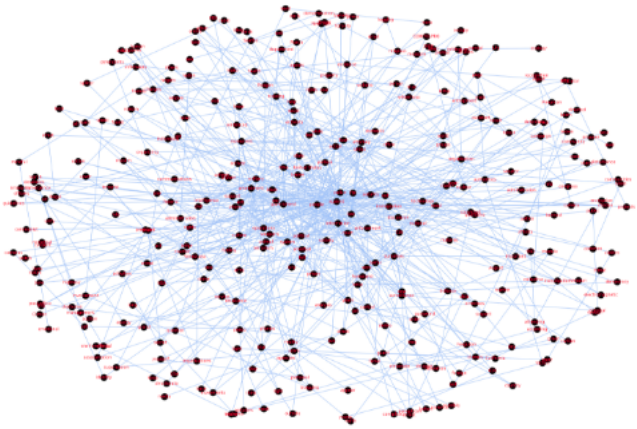

Figure 5: The original network based upon relative frequency within the term document matrices. This network is the composition or union of each document network.

Network properties of interest include graph density, assortativity, clustering, and centrality. Graph density is the ratio of the number of edges in a network over the total number of possible edges between all pairs of nodes. Let n be the number of nodes and m the number of edges in a graph. The density of an undirected graph is defined as:

$$d = \frac{2m}{n(n-1)} \tag{5}$$

Density is an analog for how related the terms are within all of the narratives. Given that the density is less than 2%, this implies that there is logic behind the discourse, or that the resulting edges between terms within these narratives are distinct and nonrandom. The density provides a foundational structure to frame the content of the narratives.

Assortativity measures how hierarchical a graph is; whether nodes of high or low degree are connected to other nodes of a similiar degree. Degree is the number of edges incident on a particular node. The assortativity of a network, r, is defined as:

$$r = \frac{1}{\sigma_q^2} \sum_{jk} (jke_(jk) - \mu_q^2) \tag{6}$$

Where $e_(jk)$ is the joint probability distribution of the excess degrees of the two nodes [nodes j and k] at either end of a random chosen link (Gnana et al). μ and σ are the mean and the standard deviation of the excess degree distribution q_k. The frequency network is disassortative (-0.0481) and shows that there is no strong correlation for high degree terms being connected to other terms of the same degree. Disassortativity shows that some of the most central nodes, or hubs, are more directly connected to peripheral terms, illustrating that these hubs are more directly apart of a larger range of documents. These hubs are not isolated and are keystones for defining the narrative in a broad way.

Clustering can be defined locally or globally. The average clustering coefficient is global, and is defined as follows:

$$C_x = \frac{1}{n} \sum_{v \epsilon G} (c_v) \tag{7}$$

Where n is the number of nodes, v is an individual node, and G is the graph. The average clustering coefficient is low, indicating that the terms

are not well cliqued, or are not within well defined neighborhoods. This can imply that conversations within the narratives are not isolated and share many common terms or concepts.

Centrality discerns the most critical vertices that define indirect connections for many of the terms. Degree centrality is the most basic type of centrality measurement. Degree centrality measures the degree of a node, which is defined as the edges incident upon the node. The vertices with high degree centrality are as follows: police(47), people(40), government(22), information(21), political(20), law(19), military(18), security(18), court(17), war (16), national(15), action(12), system(12), prison(11), federal(10), and support(10). These terms display actors and themes within the documents. Actors include "police", "people", "government", and "military"; themes include "law", "security", "court", and "prison".

To understand more about the relationships between the high degree vertices, a second network was created with its respective latent semantic information. Edges were defined between high degree terms its most similar terms found through LSA. Edge weight is defined as the cosine similarity measure between the LSA terms and the original high degree term.

Number of vertices:	235
Number of edges:	252
Average degree:	2.1447
Graph density:	0.0092
Average clustering:	0.0318
Transitivity:	0.0127
Assortativity:	-0.9168

Figure 6: High Degree Network with LSA Properties

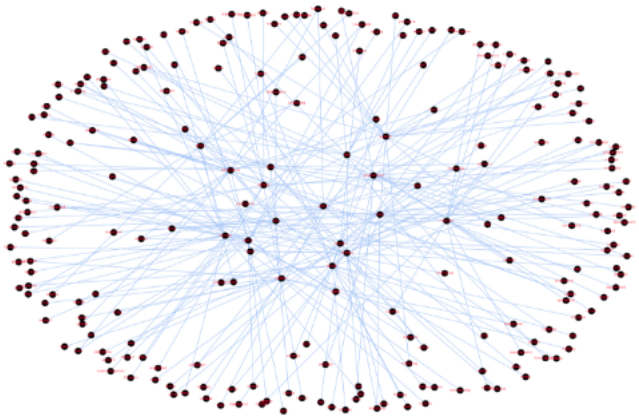

Figure 7: The second network consisting of high degree terms of the frequency network and their respective latent semantic information.

The second network is more disassortative, less dense, and less clustered, emphasizing the properties that the original network had. Degree centrality remained relatively the same (given how the second network was constructed), so alternative centrality methods were considered. Centrality measurements taken include betweenness and closeness centrality. Betweenness centrality of a node, v, is the sum of the fraction of all-pairs shortest paths that pass through a vertex, v (Hagber et al). Values of betweenness are normalized by its maximum possible value.

$$c_B(v) = \sum_{s,t \epsilon V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{8}$$

Closeness centrality is the distance of a node to all other nodes in a network, or in the case that the graph is not connected, to all other nodes in the connected component containing that node (Hagber et al). The closeness centrality of a node, u, to another node, v, is defined as follows:

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)} \tag{9}$$

Betweenness Centrality	≤ 0.15 :	: "provide", "security", "war"
	≤ 0.23 :	: "court", "federal", "law", "national", "serve"
	≤ 0.30:	: "government"
Closeness Centrality	≤ 0.13 :	: "arrested", "deadly", "force"
	≤ 0.15 :	: "convictions", "crime", "custody", "detained"
	≤ 0.17:	: "court"

Figure 8: Centrality measurements of High Degree LSA network

Themes and consequential terms are shown to be more central. Thematic terms that previously appeared ("security", "war", "court") and new thematic terms ("arrested", "detained", "crime") indicate that these are either consequences of activism or motivations for activism, given their content and juxtaposition. Actors, such as "government", appear again in the betweenness centrality measure. Dramatic and negative terms, such as "deadly" and "force" appear. These centrality measurements provide more context and indicate conflict and perceived threats, such as prosecution, incarceration, or bodily harm.

A union of the the "police" and "people" vertices with their respective LSA information alone shows that the two graphs intersect at only two LSA vertices: "versus" and "street". This illustrates a clear opposition, potentially occurring primarily at public street demonstrations.

Targeted attacks were performed upon the network to gauge whether the network remained globally well connected. Removal of central vertices indicates the robustness of a network. Central nodes were chosen by their degree centrality. If a network does not shift by a relatively high global degree, the network is considered robust. If the converse occurs, the network is not considered robust. Removal of central nodes within the high degree LSA network shows that the network is relatively robust, but that the vertex "security" shifts the network by the largest global degree (difference of 0.1532).

Node removed		
	"court":	: 2.0256
	"federal":	: 2.0171
	"law":	: 2.0256
	"people":	: 2.0171
	"police":	: 2.0000
	"security":	: 1.9915
	"war":	: 2.0427

Figure 9: Targeted attacks of high degree LSA network. Left column corresponds to vertex removed and right column corresponds to global degree of network.

The term "security" affects the network structure the most and is within all centrality measurements. The notion of security exposes itself as more significant in understanding conversations of state repression within the data set. Context to this term is provided through the line graph of the high degree LSA network. The line graph has a node in place of each prominent edge in the previous high degree LSA network, so now the vertices are pairs of terms with prominent edges, with a new edge joining those nodes if the two edges in the original network shared a common node.

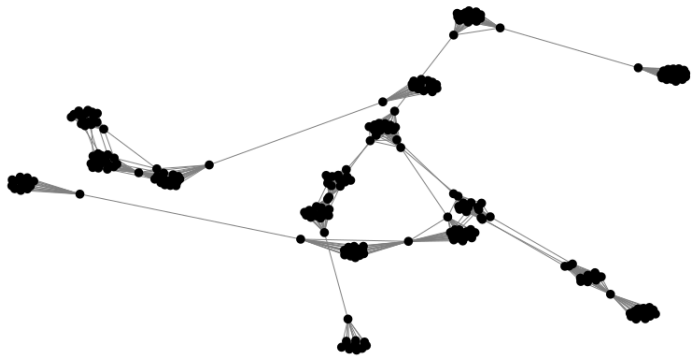

Figure 10: The line graph of the high degree LSA network.

Degree Centrality	≥ 0.07 : :	("information","police")
	≥ 0.13 : :	("data","security"),("messages", "security")

Figure 11: Centrality measurements of line graph of High Degree LSA network

The context of security becomes more evident through analyzing all network structures. The term, security, is primarily used to suggest communication privacy between activists, potential surveillance, and data collection of activists. This proves to be the one of most most significant perceived threats to experiencing state repression from activism, and this is affirmed throughout excerpts of articles within the dataset.

CONCLUSION

The most central terms and edges illustrate a narrative of conflict. Layers of opposition appear between activists within the given sample, revolving around terms such as police, security, and consequences, such as arrest. Security concerns are evident and are indicated in regards to communication privacy (i.e., high significance of vertices within line graph ("data", "security"), ("message","security")). These security concerns could be interpreted as potential compromise of private communications between activists for surveillance purposes. These interpretations are affirmed through interviews that Reynolds-Stenson conducted.

Consequential terms such as arrested, convictions, court, law, and rights had high centrality measures relative to the rest of the network. The networks are highly disassortative, showing non hierarchical relationships between minimally connected terms to central terms. This indicated that central terms are far reaching and more directly connected to peripheral topics, implicating that the central terms are keystones for defining the narratives throughout most of the data. Minimal clusering occurred within all networks, indicating that the terms were not within well defined neighborhoods, conveying a level of unification between discourse across the documents within the dataset.

This narrative of opposition parallels Reynolds-Stensons observations during her interviews. Many activists expressed sentiments that echoed this feeling of enmity seen in the documents. As such, while our analysis is not all-encompassing enough to make conclusions with any level of certainty, it may suggest that activists who have experienced state repression perceive activism as an oppositionary endeavor. This may be a cause or result of their experiences. Larger trends across all types of activism can be anaylzed with a more generalized sample of documents that span a wider array of social movements. These hypotheses could be tested with further research given a more expansive data set, but our research provides effective insight into how state repression can be perceived by activists within Arizona in recent years.

ACKNOWLEDGEMENTS

The authors thank Heidi Reynolds-Stenson for her contribution in providing our data set and for her support. We would also like to thank Dr. Ildar Gabitov and William Lippitt for their technical assistance with conducting our analysis.

REFERENCES

1. Anand, Ashish. "Complex Network Theory: An Introductory Tutorial." Department of Computer Science and Engineering. Indian Institute of Technology, Guwahati. 12 Sept. (2013).

2. Animation LSA on AP Corpus. (n.d.). Retrieved March 12, 2018, from http://topicmodels.west.uni-koblenz.de/ckling/tmt/svd$_a p.html$.

3. David Easley and Jon Kleinberg, "Positive and negative relationships," in Networks, Crowds, and Markets: Reasoning about a Highly Connected World, Cambridge University Press. (2010).

4. Fariss, C. J. Schnakenberg, K. F. Measuring Mutual Dependence between State Repressive Actions. Journal of Conflict Resolution 58, (2014).

5. Kim, K. Mathematical approach for Text Mining 1. Lecture presented in Ulsan National Institute of Science and Technology, Ulsan. (2018).

6. M.E.J. Newman, Assortative mixing in networks. Phys. Rev. Lett. 89, 208701 (2002)

7. Noldus, R. Mieghem, P. V. Assortativity in Complex Networks. Assortativity Survey (2015).

8. Rosario, B. Latent Semantic Indexing: An Overview (Rep.). (2000).

9. Saito, N. MAT 167: Applied Linear Algebra, Lecture 22: Text Mining. (2012).

10. Thedchanamoorthy, G., Piraveenan, M., Kasthuriratna, D., Senanayake, U. (2014). Node assortativity in complex networks: An alternative approach. Procedia Computer Science, 29. (2018).

11. Turner, S. Success in Social Movements: Looking at Constitutional-Based Demands to Determine the Potential Success of Social Movements. (2013).

12. Hagber, A., Schult, D., Swart, P. July 05, 2017. NetworkX Reference Release 1.11. (2015).