Modeling the Language of Risk in Activism
Initial Project Proposal

Evan Brown and Hailey Reeves
Department of Mathematics, University of Arizona,

March 20th, 2018

Keywords:
Sociology, dissent, activism, textual analysis, latent semantic indexing, network analysis.

ABSTRACT

Perceived risks associated with activism can implicate what patterns of repression have existed over an interval of time, specifically the risks associated with engaging in political dissent. Our goal is to analyze literature that documents different accounts from activists within the United States, using a data sample of texts that span across multiple decades and social movements. Our data set was provided by University of Arizona, Department of Sociology PhD candidate, Heidi Reynolds-Stenson. To create our model, we will use query matching, latent semantic indexing, and network theory. Our goal is to find underlying structures that may be used by sociologists to create further hypotheses.

INTRODUCTION

Our model will determine common perceived risks associated with activism through studying narratives from activists, specifically focusing on the risks associated with dissent. Through performing textual analysis, we will measure the most significant influences that discourage activist participation, focussing on deterrents such as state 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 are directly working with Heidi Reynolds-Stenson, a PhD candidate within University of Arizona's Department of Sociology. 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 accounts of disengagement from social movements and from experience with state repression, where the authors document recovery from or mitigation of state backlash. Examples of literature include accounts from the Civil Rights Era, where activists document their experience with covert and at times illegal surveillance (i.e., the Counter Intelligence Program or COINTELPRO), to contemporary accounts from demonstrations such as the G20 Summit protests. Our model will be created through two main methods, which are: 1) query matching, potentially employing latent semantic indexing, and 2) network analysis.

METHODS

i. Text Mining and Query Matching

In order to create models based on our compendium of texts, we must first use some form of text mining - a blanket term for any process which extracts useful information from a text. Our first step will be to perform 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 will be singular terms. To extract these tokens, we will use a program or web application to create a list of terms that appear in each document, along with the frequency with which each term appears. Once we have our word frequency lists for each document, we will perform stemming, reducing each term to its word-stem form. We will then remove unnecessary words, such as conjunctions, articles, and prepositions; these are known as stop words.

From here, we can construct a term-document matrix using the condensed word frequency lists from all of our texts. 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.

Once we have this term-document matrix, we can perform the process of query matching. First, we introduce a query, a vector corresponding to a set of term frequencies. Our goal is to compare this query to any of the document vectors in our term-document matrix. To do so, we wish to 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 a_j and a query vector q, the angle between them, θ, can be found with:

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

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

$$q \bullet a_j = q^T a_j \quad (2)$$

The geometric definition of the dot product is:

$$q \bullet a_j = ||q|| * ||a_j|| cos(\theta) \quad (3)$$

Where θ is the angle between the two vectors. We can then set the two definitions equal to each other:

$$q^T a_j = ||q|| * ||a_j|| cos(\theta) \quad (4)$$

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

The closer this value of cos(θ) is to 1, the smaller the angle, and therefore the more similar the query is to the document vector in question. For our project, however, instead of using arbitrary queries we will use the document vectors as queries for each other. We will thus create a similarity index, succinctly displaying the similarities between all documents.

ii. Latent Semantic Indexing (LSI)

While this similarity index will do a good job capturing the relationship between documents, there are several confounding factors that reduce the accuracy of our calculations. We have not, as of yet, accounted for synonymy - when different words are used to mean the same thing - or polysemy - when one word is used for a different meaning in different contexts. These two factors complicate our method, as failing to account for them could result in seeing a relationship that is not there, or in missing one that actually is. To account for this, we can manually examine the content of each article and group terms based on concept or content. By grouping terms in this way we bypass the problems of polysemy and synonymy, and allow ourselves to explore the relationship between the deeper concepts present in the documents.

While it is possible to manually group terms into these groups, we could also utilize the process of latent semantic indexing, or LSI. LSI 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. In essence, LSI will reveal the content groups already present in our document, bypassing our need to introduce bias and create the groups ourselves.

LSI 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, LSI could provide a more accurate measure of the similarity between our documents, while also extracting relevant content groups for our terms from the texts themselves.

To accomplish this, LSI 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:

$$A_{txd} = T_{txt} S_{txt} (D_{dxt})^{T} \quad (5)$$

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. 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. By choosing k such that our k-rank approximation is very close to our original matrix, we obtain a decomposed matrix that contains a very useful organization of our terms. Within this decomposition, terms that are similar will appear close together, allowing us to see possible content groups arise from our data set.

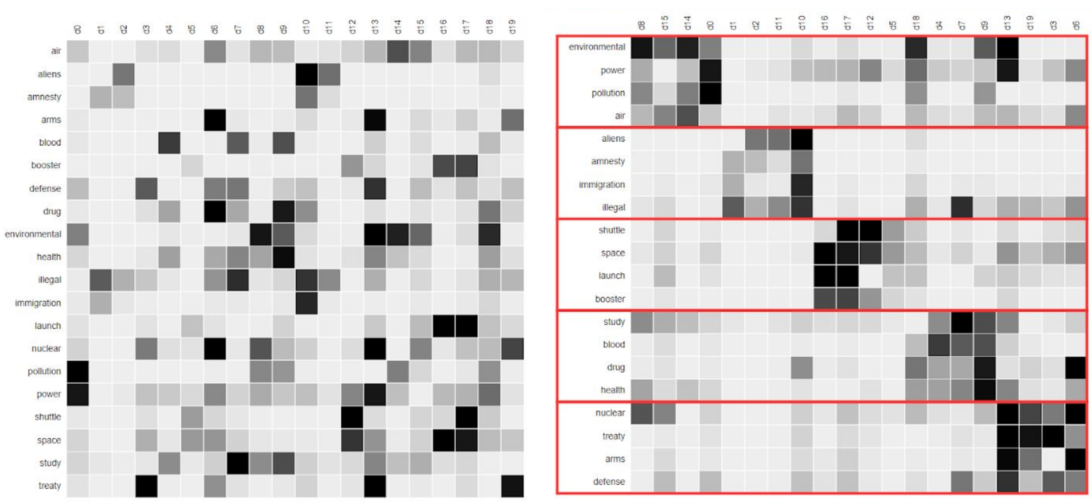

Figure 1. Before and After Latent Semantic Analysis performed on Animation LSA on AP Corpus. (n.d.). Retrieved March 12, 2018, from http://topicmodels.west.uni-koblenz.de/ckling/tmt/svd_ap.html.

In the above example, darker entries correspond to higher word frequencies. As is evident, LSI has grouped the terms into five clear content groups, outlined in red. The term-document matrix becomes organized in such a way that reveals its latent semantic information.

We are also able to perform accurate query matching using this decomposition, by adjusting a query, q, to be in the following form:

$$q = q^T X_{txk} Y_{kxk}^{-1} \quad (6)$$

Where X and Y are the k-rank approximated versions of the T and S matrices from (5).

At this time, we are not yet sure that we will be able to use LSI on our document set, as we do not know if we will have access to the correct computing and programming tools. If we are not able to use LSI, we will still be able to accomplish a similar end by manually grouping our terms. Though this will introduce more bias, it will still prove to be a satisfactory analytical method even if LSI is not used for this project.

iii. Network Analysis

From the previous method described, we will translate the content of the term-document matrices into nodes, which will be characterized by a thematic phrase or a series of related concepts/phrases through a variation of LSI. The links or edges between each node will represent the relationship between each node, so in this context, it will represent the occurrence of the nodes appearing within the same term-document matrix. The degree of each node represents the quantity of links to that specific node, where a high degree implies that it has a high number of connections, which can be used to quantify global network properties such as centrality.

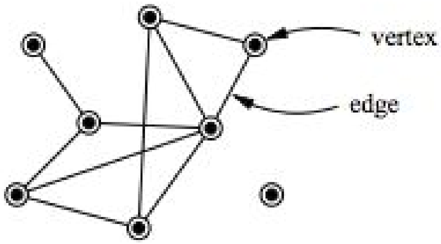

Figure 2. A small example network with eight vertices and ten edges (David Condrey, Creative Commons (CC by 2.8))

The properties of most interest to our network will be: 1) local clustering coefficients, 2) the centrality of a vertex, and 3) assortativity. The clustering coefficient can be defined as follows: let k_i be the number of neighbors of the ith node and n_i be the number of links between these neighborhood nodes, then the **clustering coefficient**, C_i is as follows:

$$C_i = \frac{2n_i}{k_i(k_i - 1)} \quad \textbf{(7)}$$

Informally, a clustering coefficient is a percentage that measures how well a group of nodes are cliqued. The point of reference is for an individual node, where a coefficient closer to 1 implies that the surrounding nodes within a loosely characterized subnetwork are connected.

Multiple measurements of centrality can be considered, but for our purposes, we will use **betweenness centrality**. Betweenness can be defined as the following: let N represent the total number of nodes within the network. Let $\sigma_l(i,j)$ represent the number of shortest paths between any two nodes, *i* and *j,* through node *l*, and let $\sigma\ (i,j)$ represent the total number of all paths between nodes *i* and *j*. Then, the betweenness centrality of the node *l*, g_l , can be defined as the following:

$$g_l = \frac{2}{(N-1)(N-2)} \sum_{i \neq j} \frac{\sigma_l(i,j)}{\sigma\,(i,j)} \qquad (8)$$

Betweenness centrality may be informally defined as the number of times a node acts as a bridge along the shortest path between two other nodes. Centrality of a vertex identifies the most central vertices, or hubs, within a graph. Centrality will help determine which perceived risks are the most influential deterrent within our data set. Our network will be weighted, meaning the edges of our network will have weights attached to them. Edge weight can be loosely defined as the strength of the connection between two nodes. The weight will be measured by aforementioned betweenness.

Removing central nodes will allow us to determine how stable the network is, or how robust it is. Removing these nodes is considered to be a “targeted attack” on a network. If the network shifts by a relatively large, global degree when removing a node, the node removed has relatively large significance in determining the overall stability of the network. If the network does not shift by a high degree when the most central nodes are attacked, the network is considered robust. In our context, robustness would indicate that the most prominent modes of repression that have appeared within the data set are not critical in determining whether or not other tactics of repression were employed or appeared within the texts.

Assortativity of a network, r, can be defined as:

$$r = \frac{1}{\sigma_q{}^2}[(\sum_{jk}(jke_{j,k}) \; - \mu_q{}^2] \quad (9)$$

Where $e_{j,k}$ 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, Piraveenan, Dharshana, and Upul, p.2541). $\mu_q{}^2$ and σ_q are the mean and the standard deviation of the excess degree distribution q_k. Loosely speaking, assortativity measures whether nodes of high degree are connected to other nodes of high degree (i.e., r = 1). If nodes with low degree are connected to other nodes with low degree, this is also considered an assortative network. If a network is disassortative, nodes are not connected in this pattern (r = 0).

EXISTING MODELS

In “*Measuring Mutual Dependence between State Repressive Actions*”, political scientists Christopher J. Fariss and Keith E. Schnakenberg created a network to derive a unidimensional measurement of mutual dependence between human rights violations. In this example, each node represented a type of human right and the edges were defined by their proximity values. In words, “the proximity value between two rights is the change in the conditional probability of observing one right violated, given the violation of another right” (Faris, Schnakenberg, 2014, p.7). The proximity values were related to the construction of this specific model. Here, the proximity between these two rights defined the weight of the edge, where the thicker edges in the network indicate a larger weight, which in turn means a larger proximity value. The edge arrows indicate the direction of the proximity relationship. The arrows do not indicate a causal relationship but a difference in the conditional probability with its converse occurring.

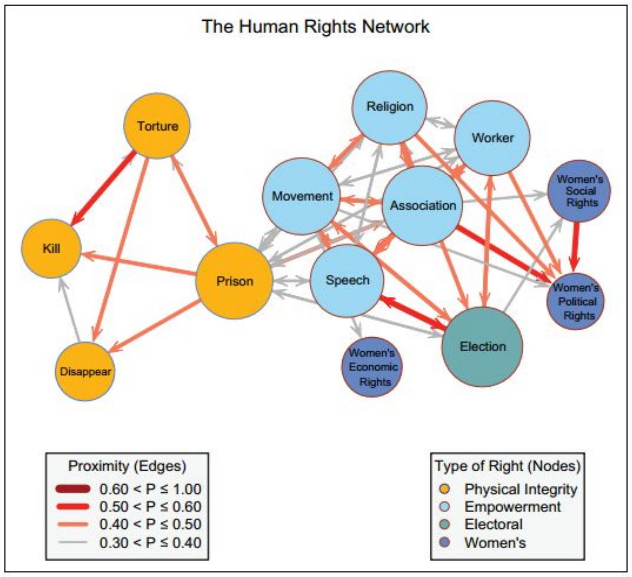

Figure 4. The human rights network, with human rights as nodes and proximity values ϕ_{ij} as edges. The plot is generated for all $\phi_{i,j,t} > 0.3$ between extreme violations in the average year. The node sizes are proportional to $\sum_j \phi_{i,j,t}$ and represent the influence of one right on all other rights in the network. The arrows should be interpreted for $i \leftarrow j$ as $P(i = 1|j = 1) - P(i = 1|j = 0)$. The arrows do not represent causal paths.

Figure 3. The Human Rights Network (from Fariss, C. J. & Schnakenberg, K. F., 2014)

TOOLS

After query matching and LSI are performed, we will use NetworkX to carry out network analysis simulations. The program is a Python based platform that allows us to analyze the node and edge attributes and perform graph manipulations, such as targeted attacks. With the program, we will be able to provide measurements for local clustering coefficients, centrality, and assortativity. Multiple models or networks will be created and will be cross-referenced to determine which networks have the most potential to draw meaningful conclusions from.

CONCLUSION

Through the two methods outlined in this paper, we will explore the relationship between the terms found in our set of documents and create interesting models displaying these relationships. By utilizing the similarity index produced from query matching in conjunction with the content groups produced either manually or with latent semantic indexing, we will be able to make educated decisions on how to manipulate our generated network. This will allow us to determine the relative importance of different terms and concepts across our set of documents, potentially allowing other researchers to generalize and formulate hypotheses on the importance of these concepts in relation to activism and state repression. We hope for our models to provide a stepping stone towards further research into activists' experiences with state repression and political disengagement, and the texts and concepts that assist them.

ACKNOWLEDGEMENTS

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

REFERENCES

1. Anand, Ashish. "Complex Network Theory: An Introductory Tutorial." Department of

Computer Science and Engineering. Indian Institute of Technology, Guwahati. 12 Sept. 2013. Lecture.

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

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

5. Kim, K. (2018, March 12). *Mathematical approach for Text Mining 1*. Lecture presented in Ulsan National Institute of Science and Technology, Ulsan.

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. Padraig Mac Carron and Ralph Kenna, Universal properties of mythological networks, EPL, 99 (2012) 28002, doi: 10.1209/0295-5075/99/28002

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

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

11. Thedchanamoorthy, G., Piraveenan, M., Kasthuriratna, D., & Senanayake, U. (2014). Node assortativity in complex networks: An alternative approach. *Procedia Computer Science, 29*. Retrieved March 19, 2018.

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