Renewable Energy Adoption in Idaho:
A Social Network Model Based Approach to Spreading the Use of Renewable Energy

Max Tanous
Bowdoin College
mctanous@bowdoin.edu
May 13, 2020

I. Introduction

The adoption of Renewable Energy to replace existing, polluting methods of energy production is imperative to slowing down the rate at which our climate is currently changing. Using Coal, Oil, and Natural Gas to produce energy is causing sever environmental degradation. There are currently many alternative options for energy production that are cleaner and significantly better for the environment. However, getting states to adopt these methods remains a challenging and important problem. The research questions this paper tries to answer, is: How do you increase the use of renewable energy in Idaho?

This paper presents a potential method for identifying specific cities to provide incentives for clean energy adoption who would then instigate the spread of clean energy adoption throughout a greater portion of the state. The method described in this paper derives from the social network problem of influence maximization. I will discuss the specifics of this problem later in the paper, but it relies on the idea that one node's (in this case a city) adoption of a new behavior is based on the number of its neighbors who currently using that behavior. A node will adopt a new behavior if the fraction of its neighbors using that behavior exceeds that nodes threshold to adoption.

This model uses my home state of Idaho as the test bed for the model. It was chosen because it currently lacks significant usage of renewable energy, but also has the land to support a variety of forms of clean-energy production. I selected 12 of the biggest cities to use in the network model.

This paper discusses Influence Maximization and Social Network Analysis, the Methods used in the model, the results from model analysis, limitations and further work to be done on this topic.

II. Influence Maximization

Social Network Analysis provides us a way to study how various scenarios will play out and effect certain connected groups of I

individuals. This first step in Social Network Analysis (SNA) is to construct and model the network you are studying. A network comprises a series of nodes connected to each other by edges. Consider a Facebook network as an example. A node would be user, and an edge would represent a 'friendship' to another user.

Once the network model is constructed, SNA is the process of examining various aspects and properties of that network. One such application of this is a problem is "Influence Maximization". The Influence Maximization problem aims at finding the most efficient way to spread the adoption of a certain behavior or product across a network. The problem statement is as follows: given a seed set size of k nodes, select the seed set that will cause the greatest spread across a network. In another words, when given a certain number of nodes k, pick a set of nodes to be initial adopters of the behavior, who will in turn cause the greatest number of other nodes to adopt this new behavior. Each node in the network has what is called a threshold. A high threshold represents a higher resistance to adopting the new behavior. Thresholds range from 0 to 1. In influence Maximization, the initial adopters have a threshold of 0, as they will automatically adopt the new behavior.

A node will adopt the new behavior if the fraction of its friends (the nodes it has edges to) exceeds its threshold value. In essence, you are trying to select the most influential nodes.

This is a very important problem in many fields. Consider, again social media. Say you are a company and you want to use social media advertisement to spread your product. You want to pick the social media users (nodes), who if you give them the product for free to advertise, will
cause the greatest number of other people in the social media network to buy your product.

I selected this as a way to answer my research question because it provides a computational method to select the optimal set of towns to incentive to switch to renewable energy to cause statewide adoption of clean-energy production.

III. Methods

This section examines the various methods for constructing the network, calculating the thresholds, determining the optimal seed set, and some additional spatial analysis work done to provide context of the specific types of renewable energy to instinctive.

A. Network Construction

The first step in the process of building my model was to select a set of cities to build the network. I settled on 13 of Idaho's largest cities, as I assumed, they would have the largest datasets associated with them. The list of cities is as follows: Boise, Meridian, Nampa, Coeur D'Alene, McCall, Lewiston. Sun Valley, Rexburg, Idaho Falls, Pocatello, Twin Falls, Hailey, and Ketchum.

To create my model, I used a Python based computer program. The program then built a network with each city representing a node. Edges were created by calculating, for each node, the three closest cities to that node. I then inputted this information into Gephi to generate a visual of the graph, as can be seen in figure 1.

This figure shows how the different cities had very different centrality measures. A node's degree is the number of edges it has connected to it. As can be seen, Boise is very central, with a degree of 8, meaning it was in the top 3 nearest neighbors for 5 other cities. This can be seen in contrast with a degree of 3, meaning it was not in any other nodes nearest neighbors list. It is important to consider degree in the context of influential nodes, as node's with high degree have the potential to influence more nodes.

B. Threshold Calculation

The next important piece in constructing the model was to identify the threshold values for each node. As stated in section II, the threshold value for a node is meant to represent their willingness to adopt the new behavior. In many Influence Maximization models these values are randomized as we often have little to know information about the majority of the nodes.

However, in my model, I wanted to use data to build these values to add to the accuracy of

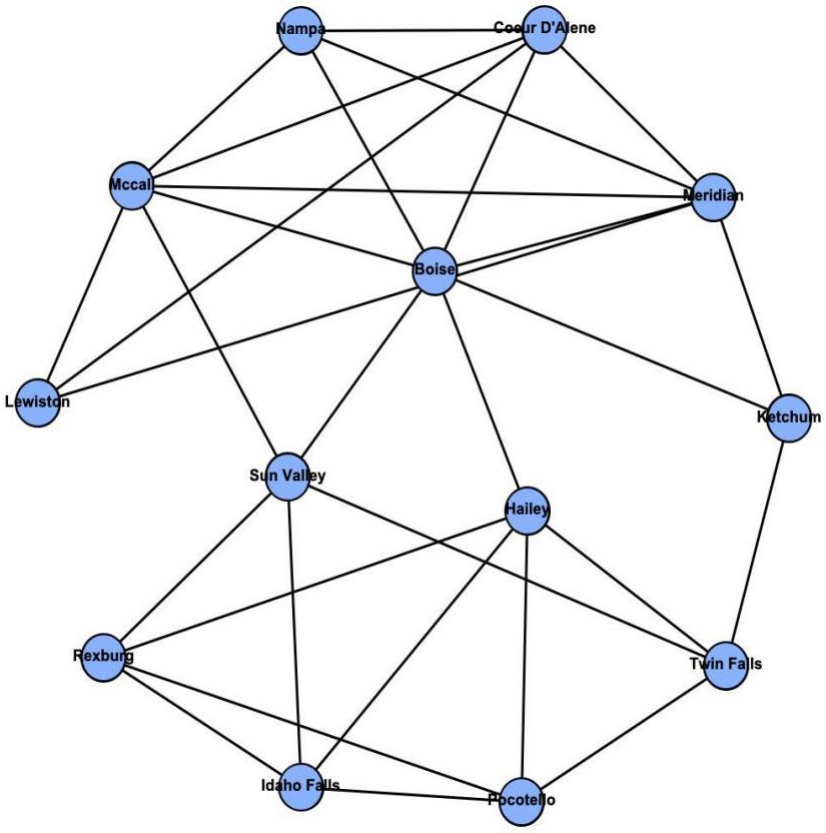

Figure 1: Gephi Graph with each city connected to its three nearest neighbors.

the model.

My model used text analysis techniques on pieces of text written about renewable energy in each city. Specifically, the python program used sentiment analysis to assign scores to each city.
Sentiment analysis is a technique that involves analyzing text based on words and phrases. It compares them to a dictionary and assigns them point values. In my program, I inputted a text document, and then cleaned it by removing common 'stop words', that are often neutral and can mislead the program. I then broke it down into separate sentences. I ran sentiment analysis on each sentence, added up their scores, and then normalized by the number sentences in the piece of text. I compared the sentence level sentiment analysis to running it over the entire document and found that the sentence level analysis was more accurate. I picked several articles and read them thoroughly so I had a sense of what their score should be, and then compared the scores to my interpretation. For each article I chose, the sentence level analysis was closer to what I determined the score to be.

I initially selected two separate datasets for comparison. I used sentiment analysis on the data sets for comparison to decide which I should ultimately end up using. The two different data sets were Articles and Legislation. For each of the datasets I used Nexis- Uni to pull the information. I collected separate sets of documents for each city.

When collecting the articles dataset, I used the following search criteria for each city:

1. Searched for "Renewable Energy Idaho"

2. Entered a sub-search for the city name.

3. Limited the Publication to only Idaho

4. Sorted by Relevance

I then skimmed through the articles to ensure they pertained to that city and renewable energy. There were varying number of results, some returning as many as 170 articles, other with only 2. If the search result for a given city was below 10, I did not save the articles and instead randomized the score for that city. Additionally, I limited the number of articles per city to 50 to ensure the data was accurate and consistent.
For each city, I ran sentiment analysis on each article, added all the scores, and normalized by dividing the score by the total number of articles, resulting in a core of 0 to 1 for each city. I followed the same process for sentiment calculation for the legislation documents. I used the following search criteria for legislation:

1. Searched for "Renewable Energy Idaho"

2. Entered a sub-search for the city name.

3. Limited the search to U.S and Idaho Legislation.

Figure 2 shows the results from Sentiment analysis from the Article data sets. Nodes are sized based on the sentiment scores for the articles for that city. Larger nodes represent a higher overall score, and a more positive attitude toward renewable energy. Blue nodes mean the score was randomized and pink means they were calculated. The total corpus size for the Article dataset was 250 articles.

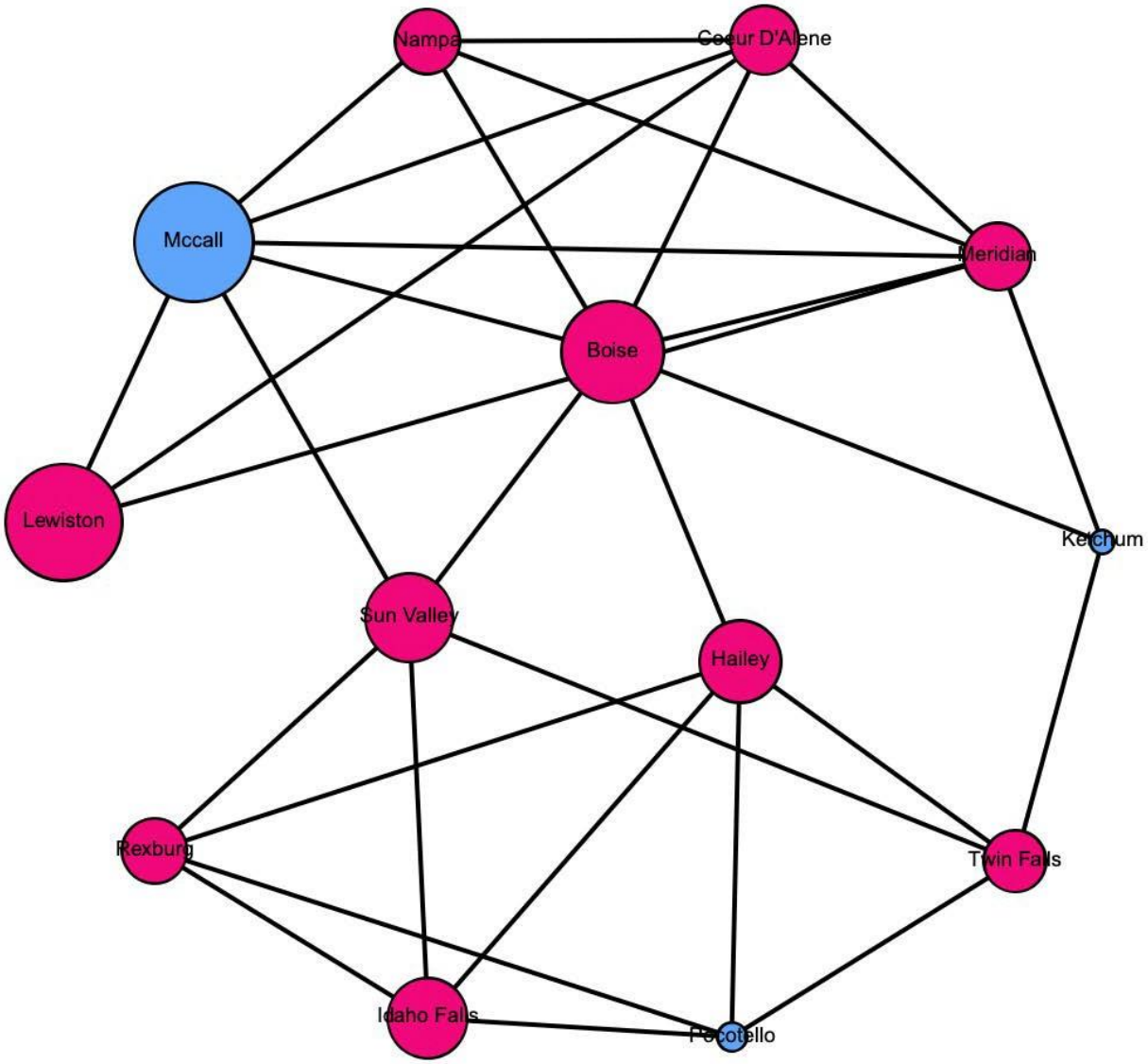

Figure 2: This network was constructed using both proximity and scores from sentiment analysis. Each node was sized according to score derived from sentiment analysis run on a corpus of newspaper articles written in that city regarding Renewable Energy. The larger the node the more positively the articles for that city discussed renewable energy. Each node was connected to its three nearest nodes.

Figure 3 shows the results from Sentiment analysis from the Legislation data sets. Nodes are sized based on the sentiment scores for the articles for that city. Larger nodes represent a higher overall score, and a more positive attitude toward renewable energy. Blue nodes mean the score was randomized and pink means they were calculated. The total corpus size for the Legislation dataset was 103 total documents.

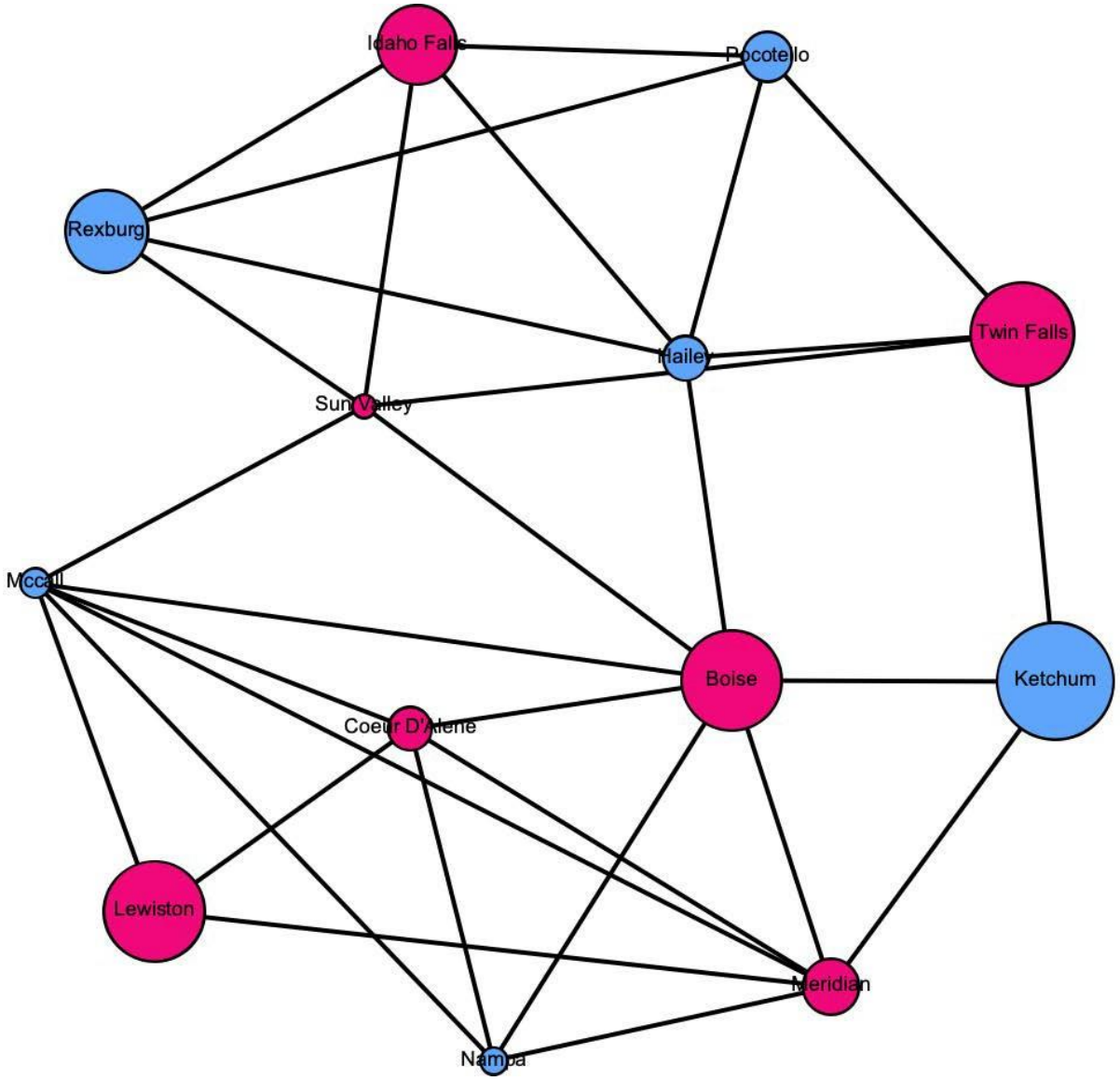

Figure 3: This network was constructed using both proximity and scores from sentiment analysis. Each node was sized according to score derived from sentiment analysis run on a corpus of Legislation regarding the city and Renewable Energy. The larger the node the more positively the articles for that city discussed renewable energy. Each node was connected to its three nearest nodes

As can be seen in the figures, I was able to collect much more data for the Article dataset. In the Article dataset I only had to randomize scores for 3 of the thirteen cities. In the Legislation dataset, I had to randomize 6 cities. Additionally, the sentiment analysis dictionary I was using in my python program did not include any legal terms, and as such the outputted scores appeared to be less accurate.

Based on this comparison I decided to use the Article dataset as my corpus. Figure 2 represents the network model that was used to calculate the maximum influence. Figure 4 shows the sentiment values for each city.

City Name	Sentiment Score
Boise	0.2645
Sun Valley	0.2292
Meridian	0.1749
Nampa	0.1678
Mccall	0.308
Hailey	0.2117
Ketchum	0.0618
Coeur D'Alene	0.1786
Twin Falls	0.1609
Pocotello	0.0751
Idaho Falls	0.208
Rexburg	0.169
Lewiston	0.3019

Figure 4: Sentiment scores for each city as calculated by Sentiment Analysis

C. Spatial Analysis

Answering the research question "How do you increase the use of renewable energy in Idaho?", is complex and reaches far beyond what my influence model results answer. Once you have determined a city to incentivize to adopt renewable energy, you need to answer the question of what type of renewable energy to adopt. It is likely that the type of preferred renewable energy differs from city to city. As, such, I decided to include spatial analysis in order to determine exactly what types of renewable energy was preferred in each city.

I used the same python program to run through the Articles dataset for each city, calculating the average number of times renewable-energy specific words were used in each city. I chose 5 keywords to identify: 'solar', 'wind', 'hydro', 'biomass', and 'geothermal'. I again averaged the counts for each city by the number of articles. Once I had the frequency of mentions for each city, I input that data into ArcGIS to generate visuals that will be discussed in the results.

IV. Results

To calculate the most influential nodes, I first had to

convert the sentiment scores into thresholds for each city. A high sentiment score means a positive attitude toward renewable energy; however, a high threshold score means the opposite. As such, I decided to first subtract each of the sentiment scores from scores from one. Because the sentiment scores all ranged .07, too .3, the thresholds would be much higher than the average degree of the node. As such, I again subtracted 0.4 from each score to match the average degrees of the nodes. Figure 5 shows the adjusted threshold scores:

City Name	Threshold Score
Boise	0.3355
Sun Valley	0.3708
Meridian	0.4251
Nampa	0.1678
Mccall	0.4322
Hailey	0.3783
Ketchum	0.5382
Coeur D'Alene	0.4214
Twin Falls	0.431
Pocotello	0.0751
Idaho Falls	0.392
Rexburg	0.431
Lewiston	0.2981

Figure 5: Adjusted threshold scores from Sentiment values

I used an adapted Influence Maximization from Hautahi King, who is referenced at the end of the paper. I will not get into the specific details of the algorithm, but it is an implantation of a well-known model: Independent Cascade. After adapting the code to fit my model structure, and inputting the threshold values I was able to generate results for different seed set sizes.

The smallest set size that could cause all the nodes in the graph to adopt the new behavior was 4. The most influential cities, based on my model were: Lewiston, Meridian, Nampa, and Boise.

Based on my model, it would be most cost effective to incentivize those four cities to adopt clean- energy production to result in the spread of adoption throughout the rest of the thirteen cities.

The next step is to determine what kind of energy to incentivize based upon the spatial analysis of the word frequency. As can be seen in the figures 6-10, the results for each city and each type of energy varied dramatically. The size of the orange nodes represents the frequency per article. Pay attention to the scales for each figure to see the differences between frequencies. We will look to examine our four target cities, starting with Lewiston. As can be seen in figures 6 and 7, Wind and Solar were mentioned with similar frequency, and either could provide an alternative to current energy sources.

The next city, Meridian, had the highest number of mentions to Hydro power relative to other cities. However, in the set of articles the most frequently mentioned energy source was solar.

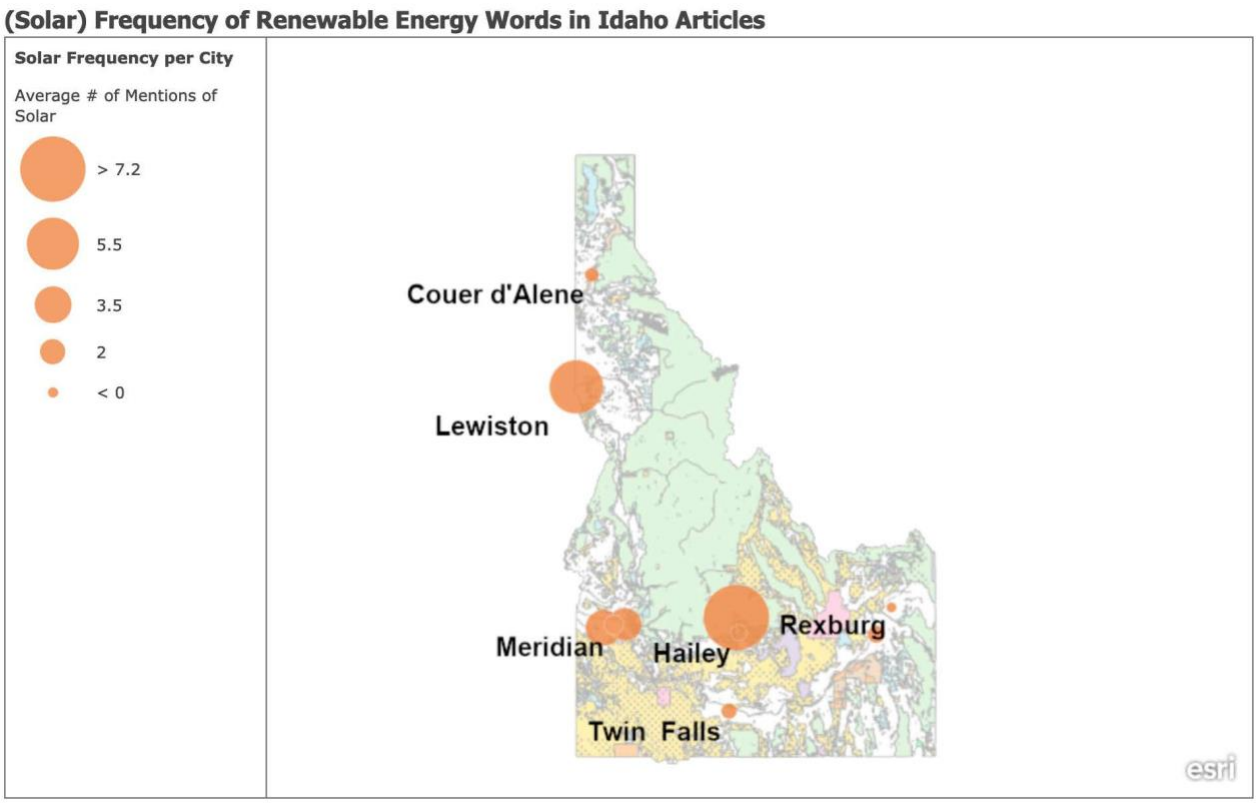

Figure 6: Spatial Representations of the average number of times the word "solar" was mentioned in each cities set of articles. This figure was generated with ArcGIS. Labels for each city are to the bottom left. If labels overlap, the city with the highest values' label is displayed

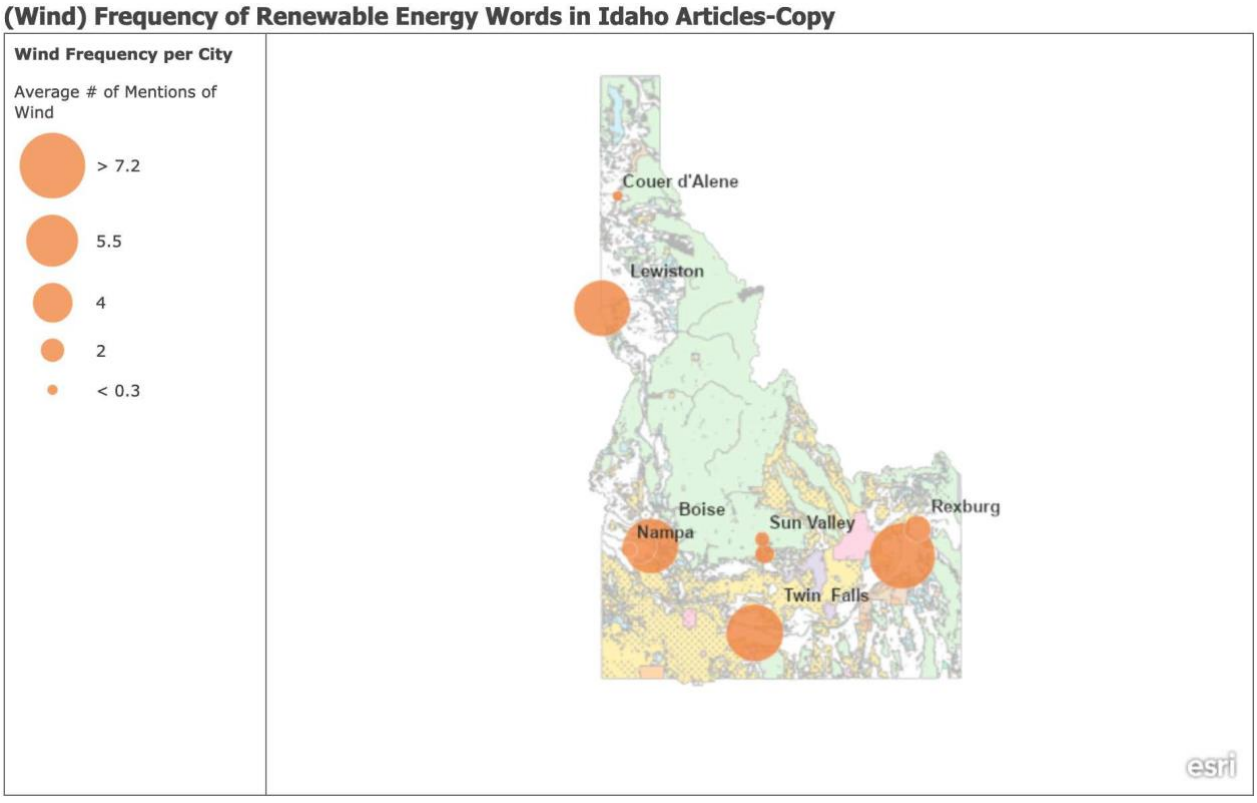

Figure 7: Spatial Representations of the average number of times the word "wind" was mentioned in each cities set of articles. This figure was generated with ArcGIS. Labels for each city are to the top right. If labels overlap, the city with the highest values' label is displayed

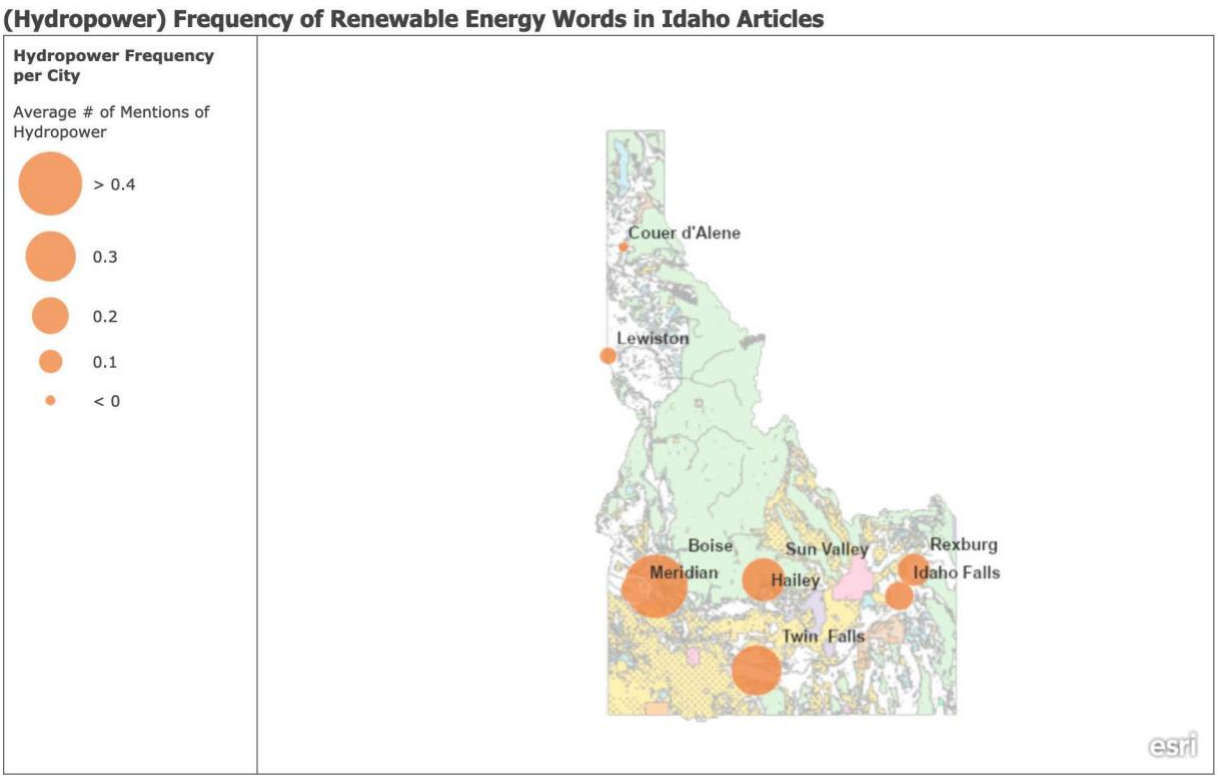

Figure 8: Spatial Representations of the average number of times the word "hydro" was mentioned in each cities set of articles. This figure was generated with ArcGIS. Labels for each city are to the top right. If labels overlap, the city with the highest values' label is displayed

---

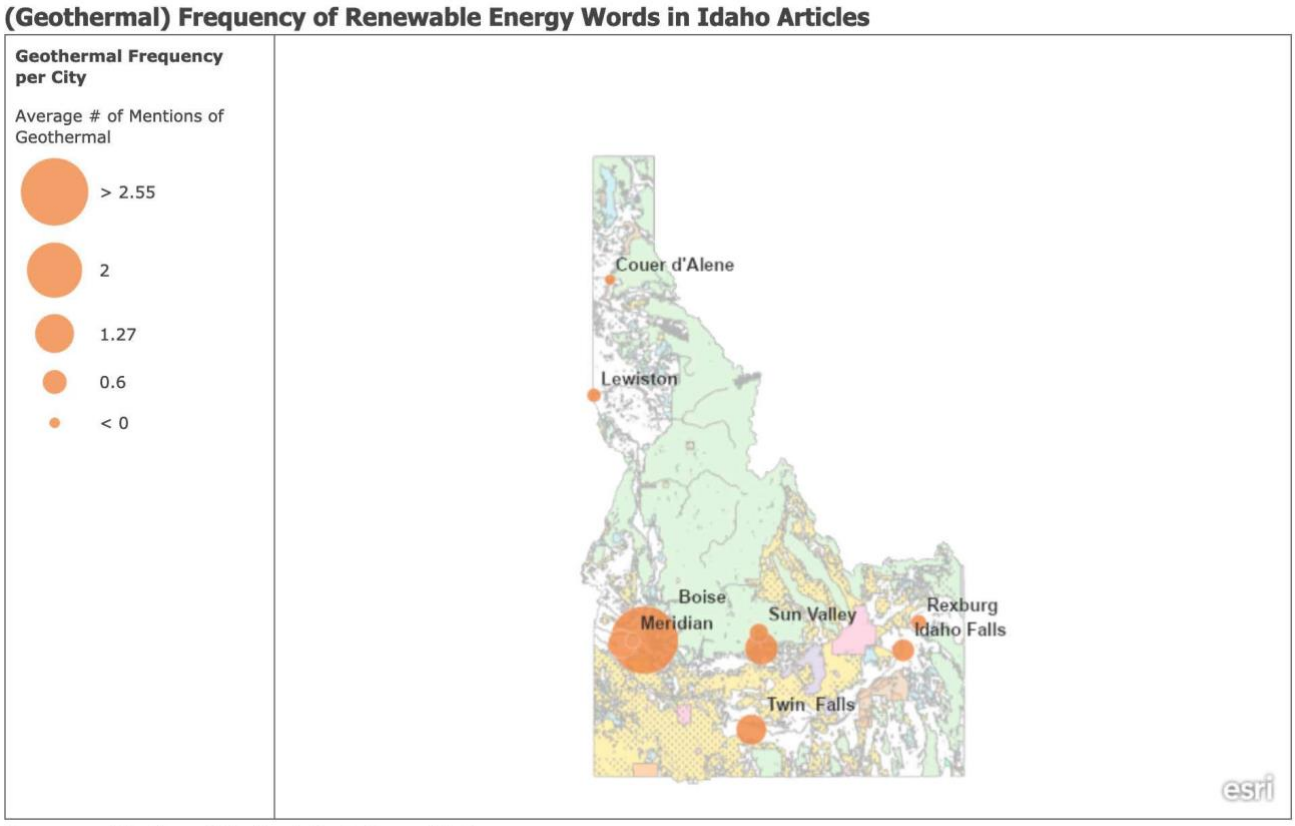

Figure 9: Spatial Representations of the average number of times the word "geothermal" was mentioned in each cities set of articles. This figure was generated with ArcGIS. Labels for each city are to the top right. If labels overlap, the city with the highest values' label is displayed

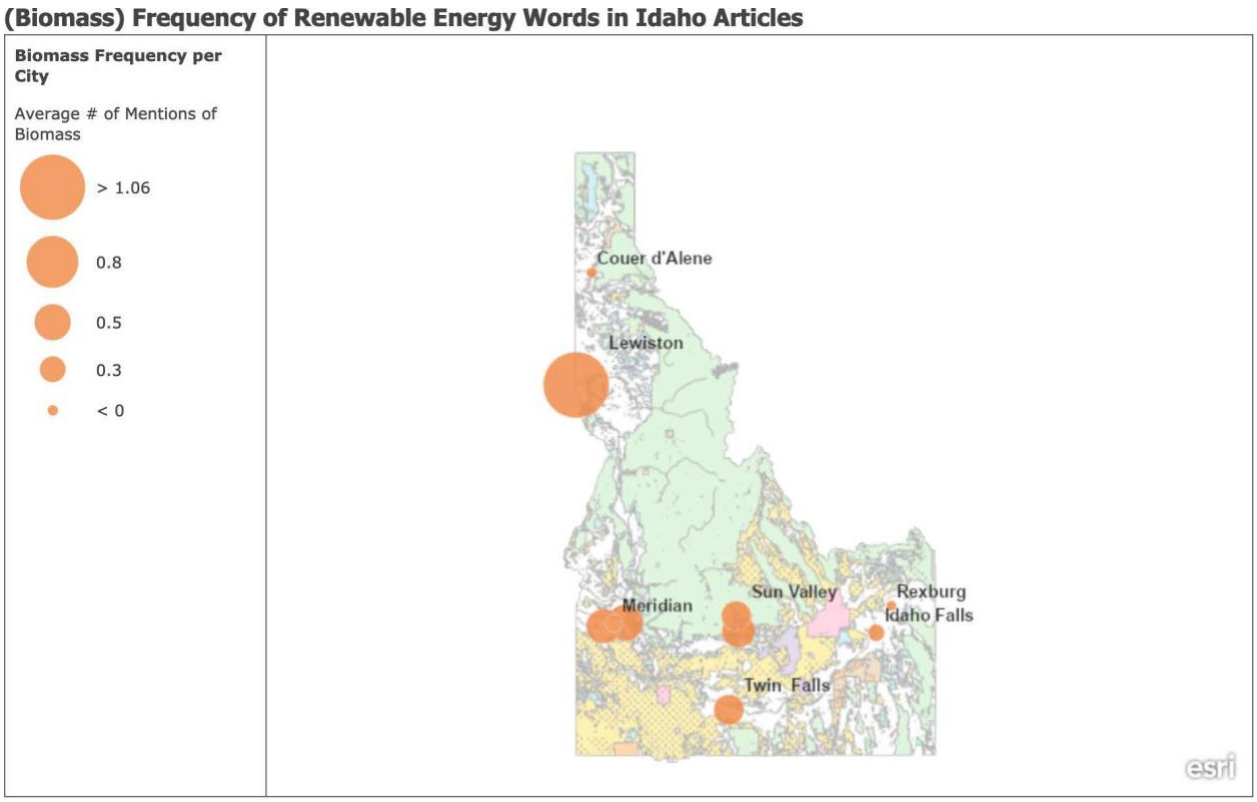

Figure 10: Spatial Representations of the average number of times the word "biomass" was mentioned in each cities set of articles. This figure was generated with ArcGIS. Labels for each city are to the top right. If labels overlap, the city with the highest values' label is displayed

Nampa had the highest number of mentions to wind in its dataset, as is seen in figure 7. In the city of Meridian, Geothermal appeared to have the highest number of mentions in its article dataset, as is seen in figure 9. Our final city, Boise, had the highest number of references to wind energy, as is seen in figure 7. At the end of the paper is a link to a more interactive map, that will allow for closer examination of each city. If cities had labels that overlapped, these figures displayed only the label for the city which had the highest value. The more comprehensive map will show more precise results.

V. Limitations and Further Work

This project had a number of limitations that resulted from time constraints. It should be used as a reference for a model which can be improved upon and applied to provide more accurate results. I will discuss potential for improvement in both the network and spatial analysis models.

A. Influence Network Model

There were several limitations in my model that could be improved to dramatically improve the overall accuracy of the model. The biggest limitation was data. A larger dataset of articles would allow for improved sentiment analysis. Having to skim each article to ensure it was relevant to my topic proved very time consuming. Additionally, for cities I was only able to find a small number of articles, extending the search from just Nexis-Uni to other databases.

The second most important piece to improve was the sentiment analysis process itself. As I mentioned, it uses a set dictionary to assign scores to specific words. However, there is no specific dictionary for environmental studies. Many important words with specific value in this field are ignored by the sentiment analysis program as they aren't present in the dictionary. My first approach to this problem was to build my own sentiment analyzer for environmental studies using machine learning. However, as is often the case with machine learning, training the model requires massive amounts of data. I was not able to collect enough data to train a model I was confident was accurate. However, if done, it would be a significant improvement to this project. Lacking a trained model, I had to use an existing sentiment analyzer. I settled on VADER, provided by Python's 'nltk' package. It had the most appropriate word dictionary. However, it was meant for social media analysis, and as such, heavily weighted scores based on capitalization and punctuation. I was able to clean the data to account for this, but the dictionary it used still lacked important words.

B. Spatial Analysis

The spatial analysis could also be improved. I checked only for word frequency and did not account for context. So, the word 'solar' could be used in the sentence 'Solar is not a viable option for Lewiston.' It would dramatically improve the spatial analysis aspect of this project. Additionally, increasing the dataset would allow for improved accuracy.

References and Resources

- Original implementation of Influence Maximization by Hautahi King: https://hautahi.com/im_greedycelf
- Project folder, including Python program code, Gephi files, Excel files, and visualizations: https://github.com/maxtanous/influence_energy_problem
- ArcGIS interactive map for word frequencies: https://bowdoincollege.maps.arcgis.com/apps/presentation/index.html?webmap=313f740e46a843dc81f22b265e35179d