Mining NIH Grant Data

Ahmed, S., Lander, S., McKenzie G., Phinney, M.

University of Missouri Columbia Computer Science

# Abstract

The distribution of research grants poses an interesting dilemma for those in the realm of academia. Many Universities and private organizations depend on funding from grants to continue their research opportunities. It would be highly advantageous for Universities and private researchers to access a broad view of how grants are dispersed as well as other functionalities involved in the process of grants being awarded. The National Institutes of Health (NIH) is a government organization that provides major research funding through grants. Since the NIH is a government entity, all of their grant history is available to the public. With the data from the NIH it is possible to find niches in the data that could be beneficial to individuals and organizations seeking grants from the NIH. With the data from the NIH it is also possible to create visual representations that can better help to understand the data. The results of the methodologies used provides both a good visualization of the distribution of grants, as well as beneficial statistics to any entity who may be interesting in requiring a government research grant.

# Background

## National Institutes of Health (NIH)

The primary source of data for the methodologies used in this paper come from the NIH. The goal of the NIH is to provide funding to research agencies who will in turn make discoveries to benefit societies overall health. Since the NIH is a government entity, the data is freely accessible to the general public. Some important data features taken from the NIH include grant information, important paper terms, award (funding amount), project start, project end, city, state, organization, etc. Without the dataset from the NIH this project would not have been possible.

## Grants

ExPORTER is a tool provided by the NIH which allows direct access to all funding data and related publications. It is free and open to everyone with a simple to use, if not always responsive, web portal. The files it provides are comma delimited, though they contain a complex quoting system due to the variety of titles and descriptions provided by PIs.

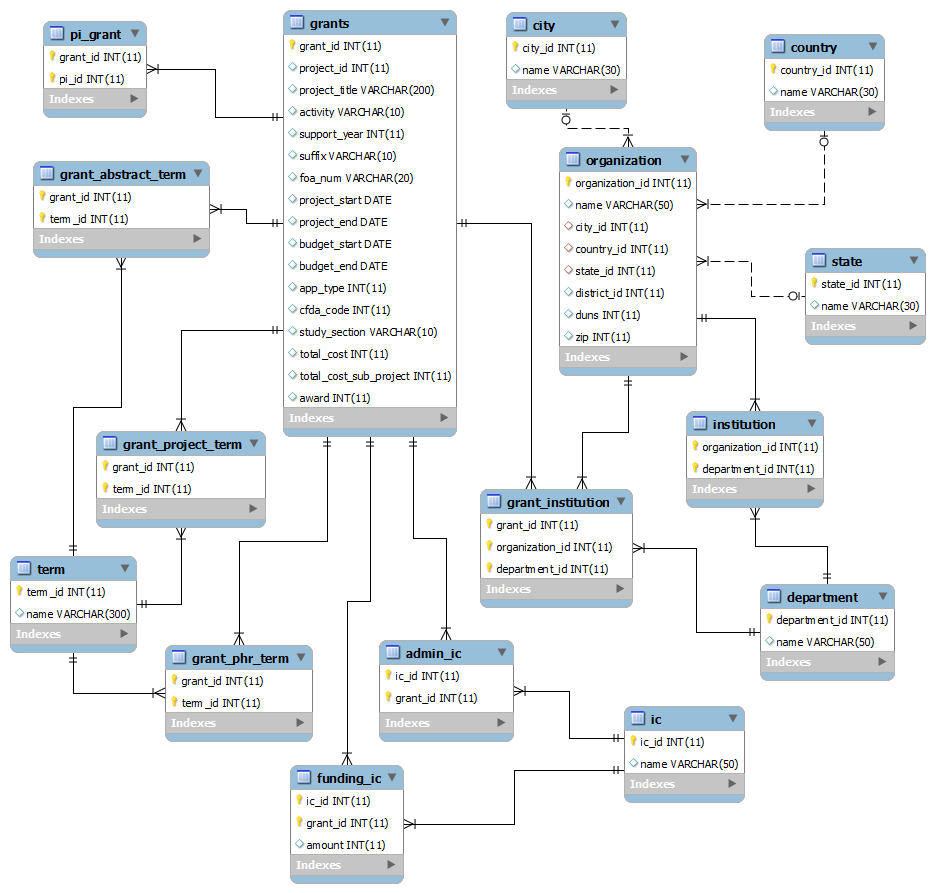
Though ExPORTER provides a large amount of information, funding information is non-existent before 2000, most likely due to institutional policies in the NIH. This causes some issues when trying to look at historical grant data, as funding is a very important part of the overall picture.

The lack of funding data wouldn't be too insurmountable a problem if not for the large amount of missing data both during this pre-2000 period as well as at the end of every year. Lack of budget date information, term information, and especially funding exists near the end of every grant cycle, making data cleanup difficult and time-consuming.

# Method

## Database Design

The raw csv data is split into smaller tables so that it can follow proper DB schema. Initially, we planned to use a MySQL database to manage our data. However, the data size soon exceeded the capacity of MySQL. As a result, we ported our data to a Hadoop[2] cluster and managed our tables with Hive[3]. Below is the ERD diagram which actually reflects the whole data.



## Co-PI network analysis

One of our goals was to understand the role collaboration plays in successful grant writing. To capture this concept we consider the co-PI network; the nodes represent primary investigators of NIH grants, and the linkages exist between nodes that have been co-PIs on a grant. We consider the following metrics to evaluate the influence of each node: betweenness centrality, closeness centrality, degree, pagerank, total number of grants received, and total amount of grant funding received. By extracting these features, we can begin to identify collaborations that are the most successful. In terms of data visualization, we decided to provide a graphical interface constructed using D3[4] to view the network.

## Geospatial Representation

As a means to view a geospatial representation of grant financial dispersal statistics, a heat map was created to display financial metrics of the NIH grant data. In order to create the heat map a template was used from the d3.js library. Once the template was established, data from both the NIH and a government Zipcode database were used to build the heat map. The heat map utilized is a visualization of every county in the United States. For instance, the more grant funding a county receives the darker shade of blue the county will appear on the heat map. The purpose of this heat map representation is to show geographical dispersion of grants across different counties in the United States.

## Grant clustering on terms

In order to get a better understanding of the terms used to describe the grants we use feature reduction for clustering. In order to do this, we utilize stacked autoencoders as proposed by Hinton[1]. Following Hinton's lead, we build a two layer neural network with a mirrored input and output, with a hidden layer size smaller than that of the input. Minimizing the reconstruction error we are able to build a set of abstracted features which can represent the input with high accuracy. These abstracted features, the activation/output of the hidden layer, can now be used as input in place of the raw input. These autoencoders can then be stacked in order to reduce the feature space to any level the user desires. By reducing the feature space to two dimensions it is possible to plot the grants in order to view possible clusters.

Before we can do this, however, we have to reduce the feature size from 117,000, the number of unique terms contained in the grant terms column. We do this by choosing only those terms which exist in at least 1 in every 100 grants, meaning 20,000 occurrences or more. This reduces our features from 117,000 to 438. After this it is just a matter of reducing the feature space from 438 to 2. Staying with the concept of a deep network, we reduce the feature space from 438 to 200, to 100, to 50, then finally 25 and 2.

# Results

## General Statistics

We calculate few general statistics using the data. When we plot few entities, we found power law distribution in almost all cases. Here grants or funds are on Y-axis and other entities like year, city, county, state, PI are on X-axis .These statistics will be shown below where left one is normal scale and right one is log-log scale except for the first one.

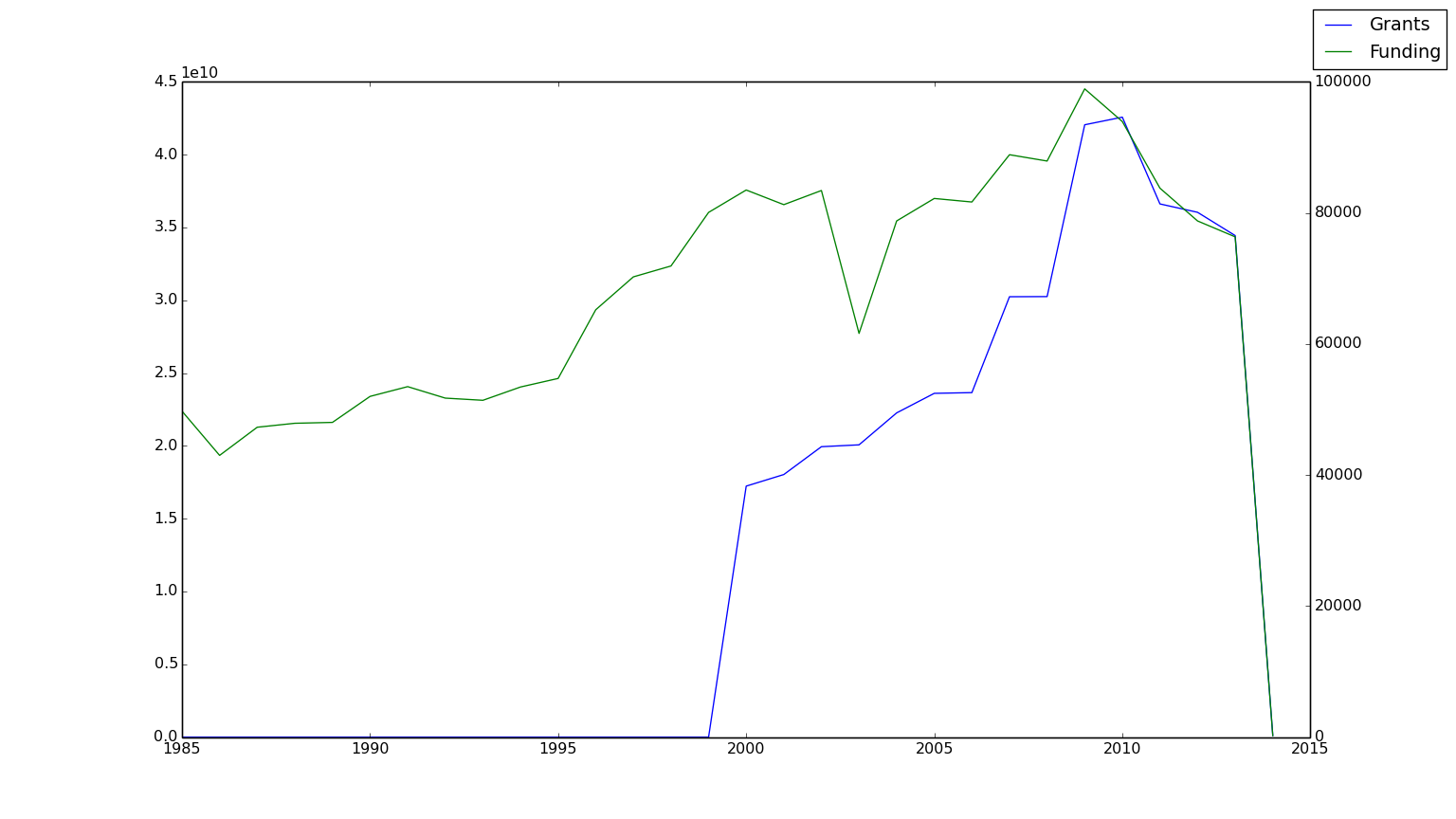


Figure 1. Grants and Funding per Year

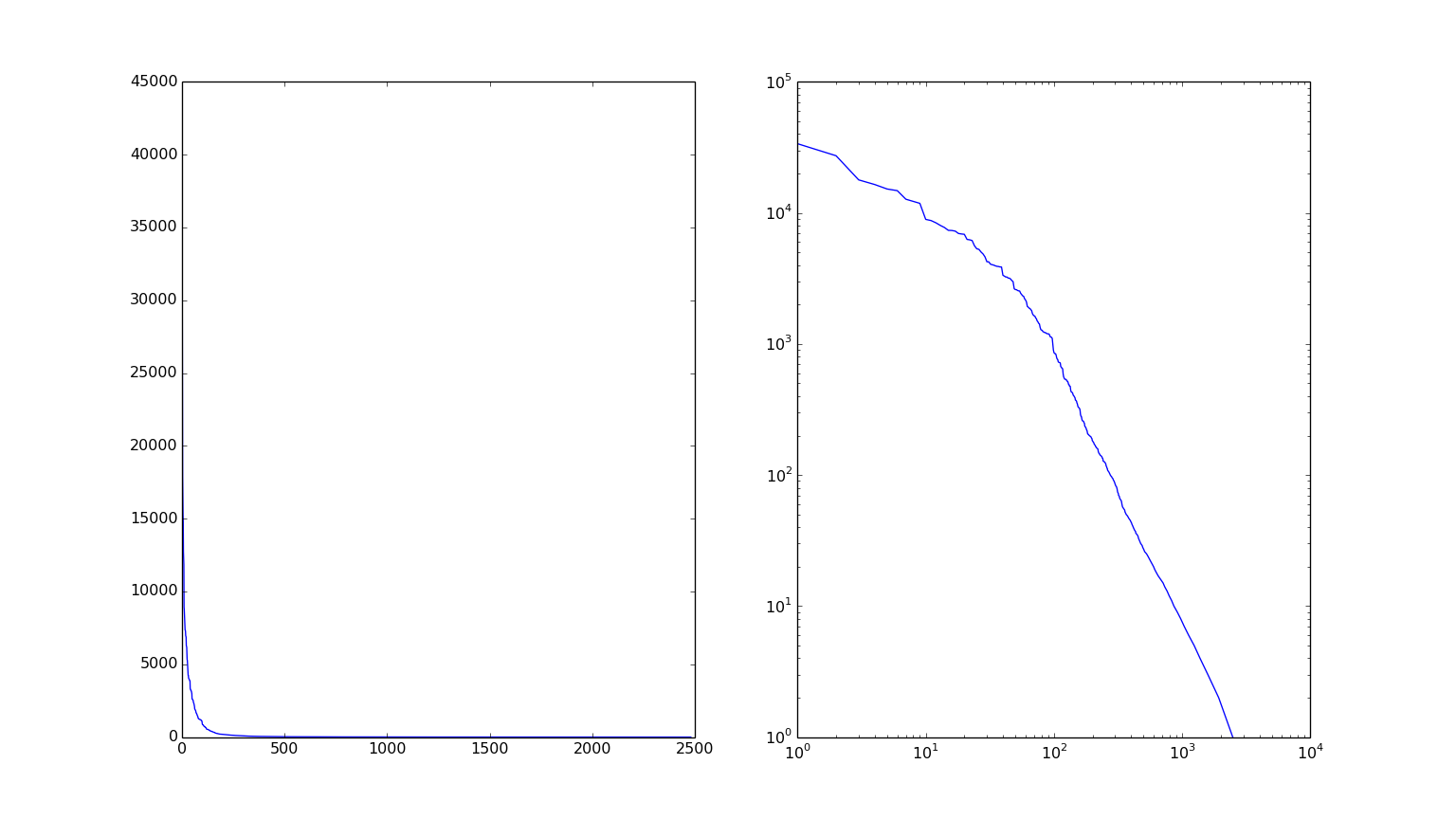


Figure 2. Grants per City

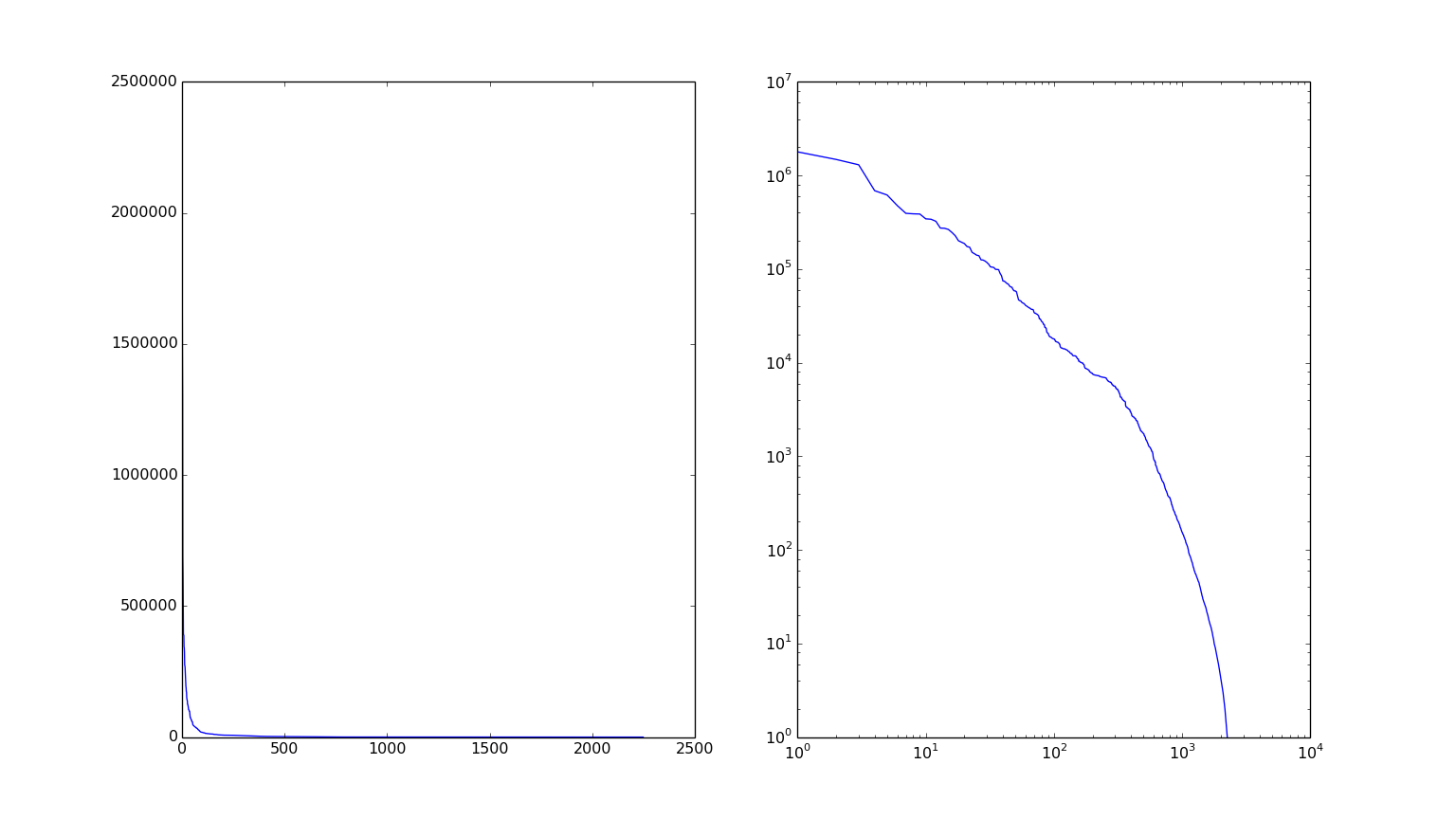


Figure 3. Grants per County

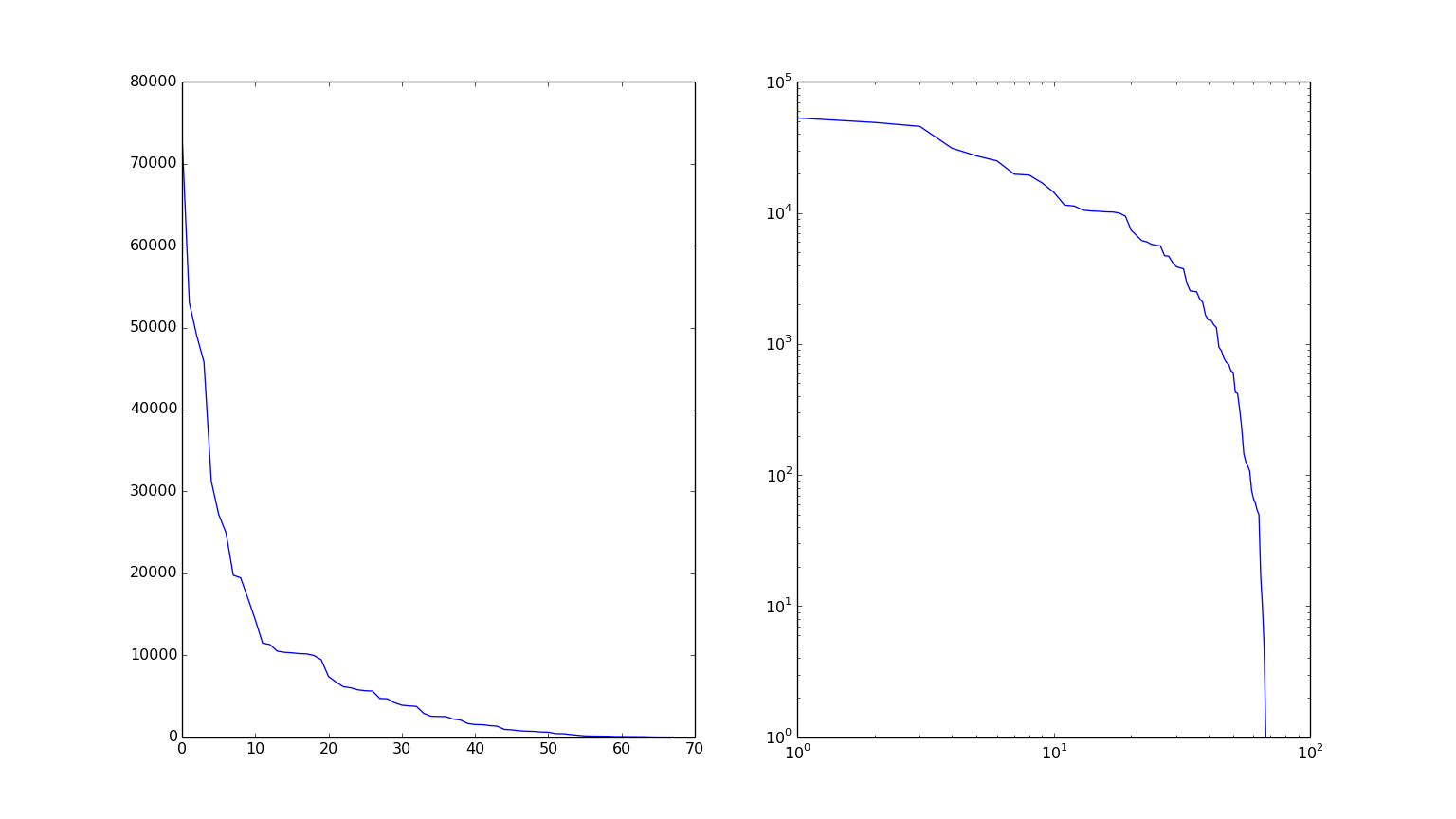


Figure 4. Grants per State

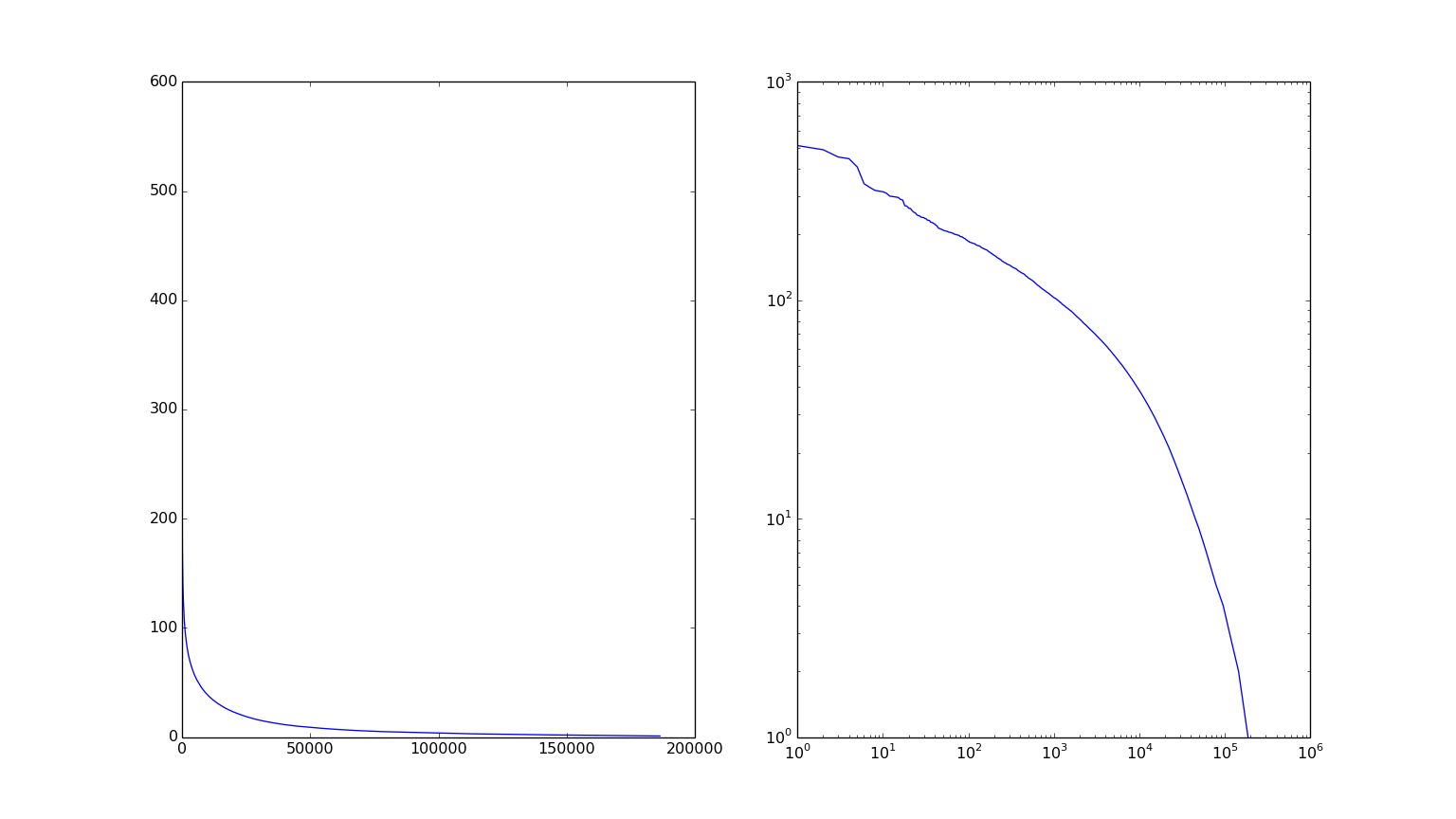


Figure 5. Grants per PI

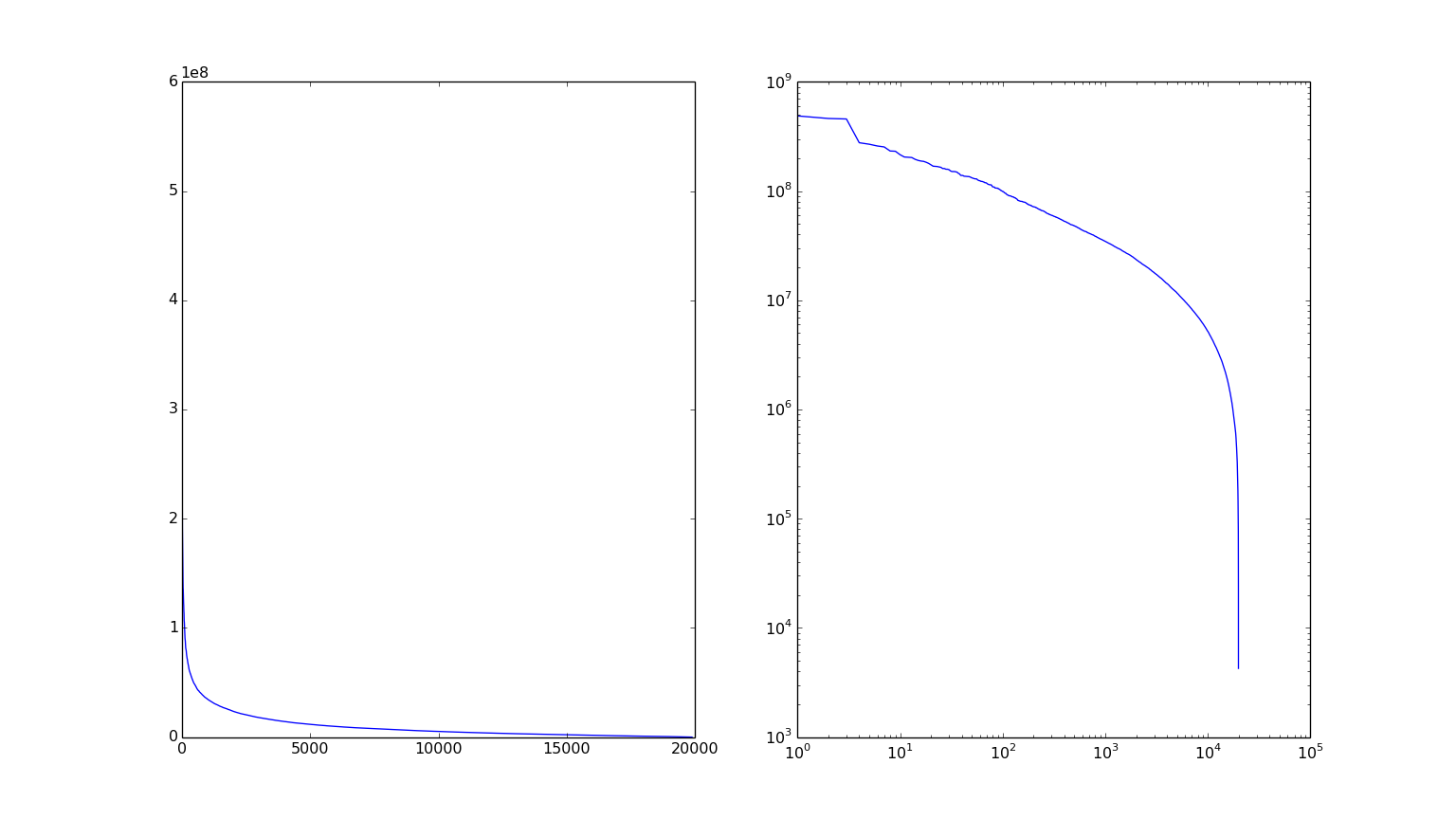


Figure 6. Funding per PI

## Network Analysis

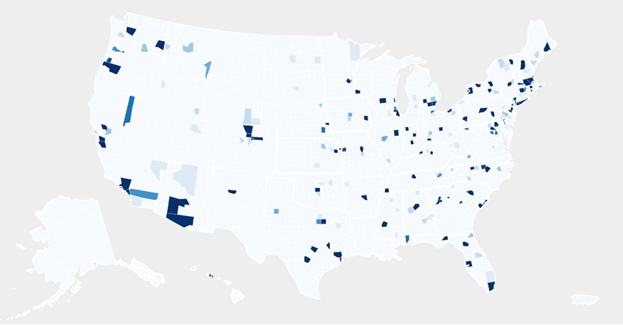
We found that the co-PI network obeys a scale-free degree distribution. In addition, we mentioned that the number of grants received and total funding received displayed power-law distributions. That is to say, most nodes are primary investigators for very few grants, and few nodes are PIs for a large number of grants. We found one node that had been the PI for 513 grants. Similarly, we found that most nodes receive little grant funding, and few nodes receive large amounts of grant funding. The most grant funding awarded to one individual across all PI roles was $565,443,151. See the figures above to visualize the distributions. To highlight a top five successful individuals, we present the NIH PI identifies below along with the number of grants and amount of funding they have received. Notice that there is no overlap between the top five listed in each category.

|  |  |
| --- | --- |
| PI ID | Number of Grants |
| 1870451 | 513 |
| 1858410 | 511 |
| 8126440 | 489 |
| 1944892 | 453 |
| 1901337 | 445 |

|  |  |
| --- | --- |
| PI ID | Total Funding Received |
| 1858798 | 565,443,151 |
| 1898145 | 488,172,834 |
| 2078516 | 462,458,526 |
| 1860045 | 457,631,046 |
| 6267091 | 277,295,411 |

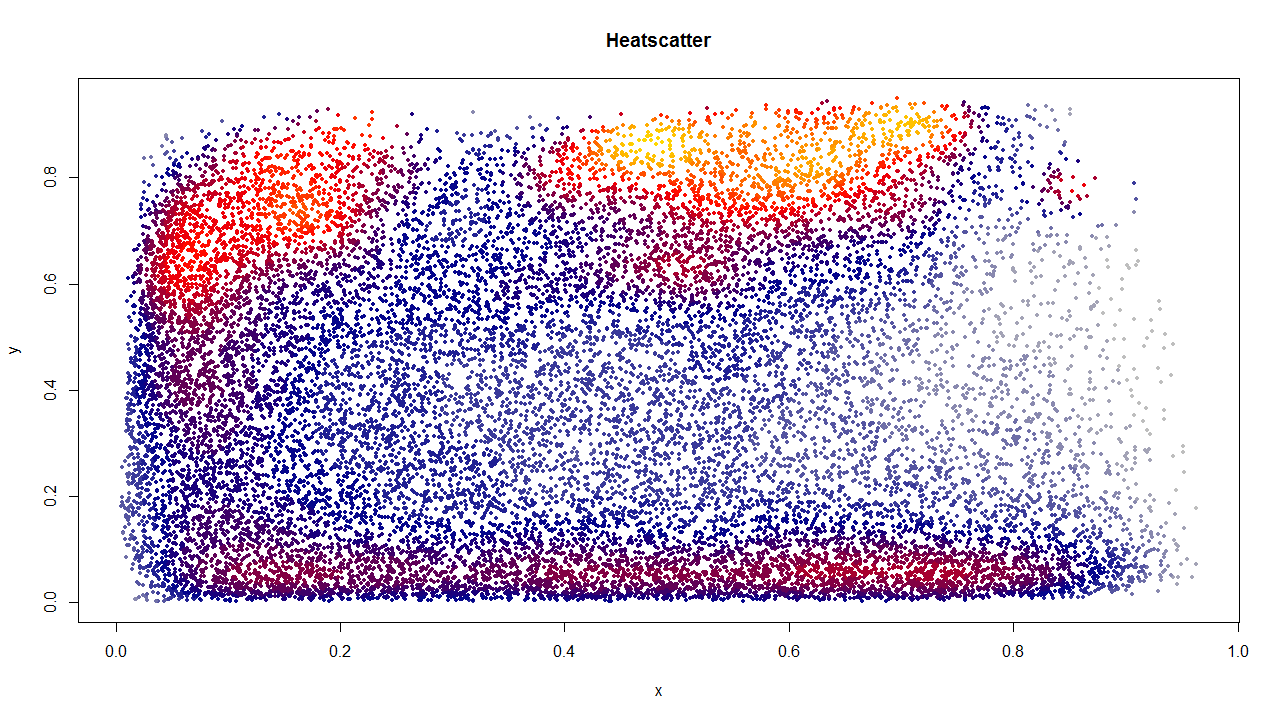
## Geospatial Representation

The image below displays the funding amounts for counties in the United States. The more grant funding a county receives, the darker shade of blue they will appear on the heat map. As you can see counties with Universities in them are often the darkest shade of blue which is to be expected. This map may be helpful in choosing which are to locate to if interested in receiving research grants.



## Grant clustering on terms

We were able to successfully reduce the feature space from 438 to 2 by using the methods described previously: 438-200-100-50-25-2. By doing this we were able to successfully reduce the dimensionality of the grant terms, visualizing it below:



This was only possible by using a distribution of 1/50th of the data, however, due to inconsistencies in the data as mentioned before. Multiple sets of data were tried with varying results, with 1/50th being the best. More analysis of the data would lead to a better understanding of why this could be the case, anything from a poor distribution of terms outside of the first 1/50th to the issues stated previously about missing and incomplete data depending on how late in the year the data is collected from.

# Conclusions

The presence of Power law distributions were prominent in the data; when we plot funding or grants per various entity we surprisingly found power law distribution in all cases. This actually reflects natural phenomena because as we know nature tends to follow this law. When it comes to terms describing the data, it is apparent that the 438 chosen are not entirely sufficient in order to do accurate clustering on the entire dataset. Expansion of the terms used may be able to counteract this issue, as would filtering of the grants before clustering. Along those lines, more preprocessing of the data could be performed as we currently use all data available with no filtering or normalization. This should yield higher quality results and better insight into the grant process as a whole.

# Future work

The NIH grant data for around last 30 years is really vast. There are lots of things that can be done using this data and a number of interesting directions should be investigated in the future. Here we propose below few future directions which may give more insight into the data. One possible future analysis of the data is looking at term importance over time, allowing us to see trends in terms which have received funding. This would be the first move toward the prediction of both term popularity as well as grant numbers and amounts based on the terms used. This information can be useful for both helping direct future grants as well as audit grants and grant committees, which is important when it comes to selecting committees as well as taking the temperature of scientific progress as a whole. Since this data spans 30 years, it would also be interesting to consider an evolutionary study on the co-PI network. There are several potential applications for link-predictive studies on this dataset.

# References

1. Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." Science 313.5786 (2006): 504-507.
2. White, Tom. Hadoop: The Definitive Guide: The Definitive Guide. O'Reilly Media, 2009.
3. Thusoo, Ashish, et al. "Hive-a petabyte scale data warehouse using hadoop." Data Engineering (ICDE), 2010 IEEE 26th International Conference on. IEEE, 2010.
4. "D3.js - Data-Driven Documents." *D3.js - Data-Driven Documents*. N.p., n.d. Web. 12 Mar. 2014. <http://d3js.org/>.
5. “United States Zip Codes.org” *Free Zip Code Database.* N.p.,n.d. Web. 14 May 1 <http://www.unitedstateszipcodes.org/zip-code-database/>
6. “U.S. Department of Health & Human Services” *National Institutes of Health*. N.p.,n.d. Web. 14 Mar. 1 <http://www.nih.gov/>