SI 671 Project Final Report

Raphael Ku

Abstract

This project analyzed vendor data collected by the Institute for Research on Innovation and Science from partner universities, converting the transactions into a network of vendors, where each node is a de-identified vendor, and each edge represents the number of unique awards and grants that utilize both vendors. Centrality measures of the network were compared to generated graphs from the Erdos-Renyi and Barabasi-Albert model to ascertain the validity of the network structure. Additionally, community detection using the Louvain modularity algorithm and Clauset-Newman-Moore algorithm was performed. Comparisons on centrality measures of the largest connected component of the vendor network versus and each large community detected using both algorithms was done, showing relatively similar clustering coefficients, with the communities showing higher Freeman degree centrality coefficients in general. Lastly, the distribution of the node attributes were extracted to see if there are patterns in vendor utilization by federal agencies or universities.

1 Introduction

In 2017, the University of Michigan reported spending $831.8M in federal funds to conduct research aligned with the goals of federal departments such as the Department of Health and Human Services, the National Science Foundation, and the Department of Energy (Service, 2017). Part of the responsibility of receiving federal funding for research is the need to report on how said funding is utilized by the Primary Investigators (PIs).

As part of the requirements of federal grants, a significant amount of data on federal funding for research and development is available, enabling meta-research on the effects of funding on individual researchers and organizations, with some focusing on collaboration through a citation network, team development in relation to funding, and innovation leading from funding. Similarly, this project focuses on doing meta-research, focusing on a research institution's interactions with vendors that supply the necessary materials and equipment to enable innovation and research, and how each vendor effectively interacts with one another in a vendor network. The analysis method chosen to do this is through network analysis, evaluating vendors through the centrality measures of the network as a whole, as well as implementing the Louvain and Clauset-Newman-Moore community detection algorithms to identify communities of vendors, and find commonalities in transaction behavior.

Due to the nature of this project, the primary audience of the results would be the partner universities featured in the dataset, specifically with staff in charge of procurement of materials pertinent to grant-funded projects and research. Additionally, the results of this project may be of interest to grant-issuing federal departments, as they may have the necessary influence to set up deals with larger suppliers/vendors to reduce research costs.

2 Problem Definition and Data

The main aim of this project to explore the interconnections of vendors utilized by research institutions. The ability to determine which vendors are related based on a research organization, or an issuing federal department may be able to help find potential improvements in the procurement process of an institution such as determining potential dependencies towards a specific vendor that may need alternatives for the process to remain robust.

The data for the project will be the UMETRICS 2017Q4A dataset provided by the Institute for Re-

search on Innovation and Science (IRIS), primarily looking at the vendor transaction files as the focus for this study. The vendor transaction files are in an SQL table containing approximately 17M transaction records, each containing information on the specific grant that provided funding, which organization was awarded a grant, and the total charged to a specific vendor, among other features.

The dataset itself is represented by 205,627 unique vendors which have at least $1000 in transactions for the past 10 years. Based on initial impressions, this number can be reduced even further by removing name duplication of corporate names, by standardizing the non-use of suffixes. Notable in this is that, in the 10 years of records available, most vendors are seldom used, whereas there is a core of 361 vendors that have been paid by more than 500 awards over the 10-year period

Unique awards	Number of vendors
1-50	202,620
51-100	1,369
101-500	1,227
500+	361

Of those 205,627 vendors, 188,685 have records indicating that the grant that provided payment also provided payments for at least 1 other vendor.

Due to the nature of the dataset as submissions from partnered universities, the analysis described here can be expanded to include new information provided by new partner universities over time.

3 Related Work

Network analysis using similar data on grant funding usage, as well as using the UMETRICS dataset focuses on the interconnectedness of research organizations, primary investigators, and team-building versus federal funding. Due to the similarity of the subject and methodology, these papers are the focus of this section. Note that no direct parallel research was found on vendor networks and federal research grant funding.

In their working paper, Lane, Owen-Smith, Rosen, and Weinberg focused on looking at researchers funded by federal grants as the focus of their paper, categorizing researchers in terms of occupational categories, such as track, training, research role, and so on. More interestingly, they performed network analysis using record level wage data to create links between individuals involved in the whole research process, including technicians, programmers, and other roles generally not cited in research publications. In line with the current problem of using vendor networks to determine relationships between vendors, this paper uses both ego networks, as well as the dataset having ground-truth labels for communities (University Affiliation) to detect outliers, such as researchers not connected to the graph. Of interest specifically would be the analysis of teams as being linked by overlapping sets of grants, similar to how vendors in the dataset are connected to one another through overlapping federal grants. The paper indicates the use of university affiliation and subject areas to build community labels, which are both present in the data as the submitter of the data, as well as the CFDA number indicating the type of grant provided. Using this framework can help me in setting up the community labels, to evaluate if there are communities in vendors at all, or just one dominating cluster of vendors for all universities and subject areas. (Lane et al., 2014)

Kardes, et al looked at researchers awarded funds from National Science Foundation grants. In terms of potential solutions offered by the paper, the organization collaboration network previously mentioned in the proposal becomes more important, as it displays attributes very similar to the vendor network, in terms of a central cluster of interconnected nodes in the graph. The paper offers an interesting option to do a historical perspective, iterating over the 10 years of data to see changes in network characteristics with the increase in unique vendors and connections over time. This may indicate that, aside from community analysis, time-series analysis could be possible with the vendor network. (Kardes et al., 2013)

For the paper by Folkstad and Hayne, while they focused on a longitudinal network analysis using citation data, their analysis was focused again on the change of graph characteristics, such as average degree, and betweenness, over time. One interesting approach that could be applied to the vendor network would be the measurement of cut-ties, or edges that can separate a component into two when removed. They note the use of this as characteristic of information flow. In terms of my paper, given that the vendor network is characterized by having many nodes that have weak ties, with a strong core of vendors that share with a large number of vendors, I can try to approach the issue of network resiliency in evaluating how

many cut-ties exist in the current vendor network. (Folkstad and Hayne, 2011).

4 Methodology

The vendor network edgelist was generated through SQL queries to the IRIS dataset, ensuring we only capture relevant vendors by requiring at least $1000 in transactions over 10 years, as well as removing all vendor names that were left blank. Additionally, an additional requirement of having more than 1 shared award in the network was added to generate a separate edgelist.

This lead to generating 2 vendor networks, where each node is a vendor, edges in said networks are weighted based on the number of unique awards shared by 2 vendors, with node characteristics for each vendor indicating which universities and federal departments they have served in the past 10 years. The larger network, composed of 188,685 nodes and 17M edges, was used to validate the use of the network structure for vendors, given that the links of the vendors more closely resemble collaboration networks, rather than social network relationships. The network was evaluated against graphs randomly generated through the Barabasi-Albert and Erdos-Renyi models, looking at the similarity and differences in number of components, average degree, and Freeman degree centrality.

The network with additional filters was comprising of 49464 nodes and 1.27M edges. Further reductions in vendors were done through only evaluating the largest connected component, removing singleton vendors, and very small communities, leaving 44774 nodes, and 1.26M edges. The Louvain modularity (Python-Louvain implementation [1]) and Clauset-Newman-Moore (Stanford Network Analysis Project implementation [2]) community detection algorithms was used on this network to generate communities in the network based on modularity maximization. The rest of the analysis focused on large communities, defined as communities with 1000 vendors or more. The average degree, Freeman degree centrality, and average clustering coefficient was extracted from each community, and compared to the largest connected component to look for differences in a community's network structure as compared to the aggregate network. Lastly, using the binary node attributes or university affiliation, and federal departments served, I looked at each community to look for interesting behaviors and patterns in communities generated.

[1] Aynaud, Thomas. *taynaud/python-louvain*. https://github.com/taynaud/python-louvain

[2] Lescovec, Jure. *Snap.py - SNAP for Python*. http://snap.stanford.edu/snappy/index.html

5 Evaluation and Results

5.1 Analysis versus Generated Networks

The initial evaluation of the network focused on general network characteristics, namely, the number of connected components, the percentage of the network covered by the largest connected component, the average degree of the graph, and the Freeman degree centrality measure. The larger vendor network was compared to graphs generated through the Erdos-Renyi model, representing random linkages between vendors, and graphs generated through the Barabasi-Albert model, representing preferential attachment to high-degree nodes. The generated models were run 20 times, with the same number of nodes as the vendor model. The Erdos-Renyi model used a p of 0.01 for node linkage probability, while the Barabasi-Albert model was set to generate 25 edges to have a similar number of edges to the vendor network. The upper and lower bounds represent a 0.99 confidence interval in the values.

5.2 Community Detection and Centrality

After evaluating the validity of network analysis on the vendor network, the Louvain modularity and Clauset-Newman-Moore (CNM) community detection algorithms were applied to the largest connected component of the filtered vendor network. This generated 120 and 440 communities respectively. Looking at only communities with 1000 vendor nodes or more, we are left with 10 communities from the Louvain algorithm, and 6 communities in the CNM algorithm.

Initial analysis of the communities focused on each subgraph/community's network characteristics as compared to the largest connected component, looking at average degree, Freeman degree centrality, and average clustering in the network. The baseline used for this analysis was the centrality measures of the largest connected component, to identify if communities have attributes distinct from the aggregate.

Based on Table 2 (in Appendix), the communities for both algorithms show that the average degree for communities are lower than the whole

Network	Ave. Degree	#Connected Component	%Connected Component	Freeman centrality
Vendor	51.40	899	94.16%	0.1766
ER Lower	494.29	1	100%	0.0016
ER Upper	495.00	1	100%	0.0022
BA Lower	49.97	1	100%	0.0318
BA Upper	49.97	1	100%	0.0420

Table 1: Results of baseline and current network

network, barring 1 community that has a higher average degree than the entire connected component, such as Louvain Community 6 displaying an average degree of 135.50 with 2,822 nodes as compared to the connected component's average degree of 56.43 with 44,774 nodes. Additionally, Freeman degree centrality is generally higher in the communities in comparison to the largest connected component, while average clustering for communities stay relatively close to the value provided by the largest connected component.

5.3 Community Detection and Service Distribution

Using node attributes on agencies and university IDs served for each vendor, each community was evaluated to detect patterns in distribution of node attributes.

Given the node attributes, it was found that the largest communities generated in both the Louvain and CNM algorithm are generally homogeneous in terms of federal agencies served. The general pattern is that, due to the sheer number of awards and grants, the National Institute of Health (NIH) and National Science Foundation (NSF) are well represented in all communities, while the United States Department of Agriculture is usually underrepresented in these large communities. Additionally, for both algorithms, 1 community effectively acted as an outlier, where 1 community in the CNM algorithm showed a stronger preference towards NIH, with 80% of vendors in the community serving grants from NIH, and 1 community in the Louvain algorithm showing an opposite reaction, serving NIH grants much less, with only around 30% of vendors in the community having provided services related to an NIH project.

Interestingly as well, for both community algorithms, we determined that the individual communities have strong identification with 1 or 2 specific universities, represented here as a de-anonymized University ID.

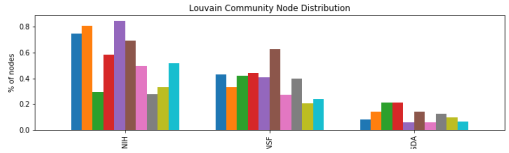

Figure 1: Louvain distribution of vendors across NIH, NSF, and USDA

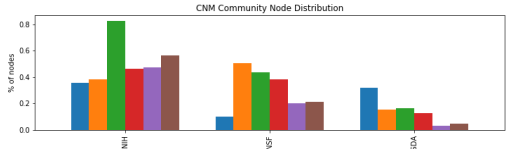

Figure 2: CNM distribution of vendors across NIH, NSF, and USDA

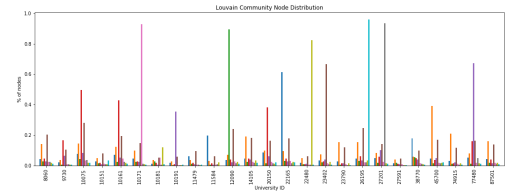

Figure 3: Louvain distribution of vendors across universities

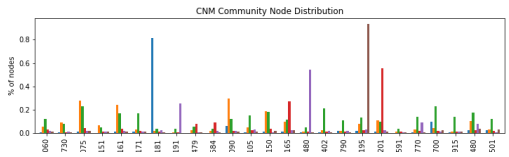

Figure 4: CNM distribution of vendors across universities

6 Discussion

Based on the initial analysis of the aggregate vendor network, the results seem to indicate the vendor network as having characteristics of preferential attachment, similar to the Barabasi-Albert graph, echoing social network structures. Additionally, the much larger Freeman degree centrality measure in the vendor network indicates a lower degree variation in the vendor network compared to the generated graphs, which may indicate high clustering and clique formation. Additionally, upon evaluating the large amount of connected components detected in the vendor network, it can be noted that the network is effectively one very large connected component, with singleton vendors and small, isolated communities in the periphery, most likely due to specific awards or name duplication in the dataset.

On the comparison of centrality measures comparing the largest connected component to communities detected in the Louvain and CNM algorithms, we found that average clustering is relatively stable across the largest connected component and communities detected, while the variation is primarily in the Freeman degree centrality measure and average degree. Since the Freeman degree centrality measure effectively is a normalized measure of the variation in degree distribution for each subgraph, we can say that, while each community is densely connected to one another, the general increase in the Freeman degree centrality seen for most communities as compared to the baseline of the largest connected component indicates that there are central or key vendor nodes in each subgraph/community.

Lastly, for the attribute distribution for each community detected, the fact that most communities have a specific preference for 1 or 2 specific universities points to the possibility that the community detection methods may be grouping vendors through attributes not present in the data provided, rather than simply a preference towards a specific university. A reasonable assumption would be that the vendors in a specific community that show a strong relationship to a university may be in close proximity to the university itself. Additionally, it should be noted that, just because a specific community has a distribution pointing to preference towards a specific university, it does not mean that the community comprises the majority of that university's vendors. Nevertheless, the distribution of node attributes on federal agencies served and universities served indicate a potential hypothesis that vendors don't have particular preferences towards agencies, or the original source of funding, but rather, toward their direct contact, namely the university or researcher paying them.

7 Things left undone

Aside from the metadata taken for each vendor on which of the 3 largest federal agencies they've served, as well as which universities they've served in the past 10 years, the data from Institute for Research on Innovation and Science contained other forms of metadata as well, such as the address of vendors themselves, the Catalog of Federal Domestic Assistance of each award. It is not unreasonable to incorporate additional metadata as node attributes, to see if the metadata can explain the communities generated

One issue with the communities generated here is that there is most likely overlapping communities present, given that a vendor can serve any number of grants, researchers, and universities, especially within the 10 year time period allotted. Because of this, the communities generated using the current algorithms don't represent the complexity of the vendors fully. Another aspect would be evaluating the communities generated themselves, in terms of quality of communities found. Ideally, more time could have been taken to look for baselines or evaluation metrics to quantitatively validate the communities found.

Lastly, one key aspect noted while working with the data is that some of the largest vendors identified are actually duplicates that were missed during the cleaning process. As the Institute for Research on Innovation and Science already does a significant amount of work cleaning vendor names, iterating through a list of a few million company names each year, significant vendors are still left as duplicates in the list due to variations in a university's encoding of a transaction. If given more time, more time will be spent on looking on how to improve the data cleaning process as a whole, which will most likely lead to a denser network with fewer nodes, and may provide a much more different result in terms of communities generated, and the network properties found.

8 Work Plan

Initial Workplan

Dataset	Description
Oct 8-19	- Run project with supervisor in IRIS - Gain access to data outside of internship
Oct 20-Nov 2	- Search for more related literature - Check SI608 slides for analysis methods
Nov 5-11	- Data preprocessing
Nov 12-18	- Initial data model
Nov 19-25	- Request review and extraction of initial results - Prepare draft of project update report
Nov 29	- Submit project update
Nov 30-Dec6	- Improvement on network analysis model - Prepare project presentation
Dec 7	-Submit project presentation
Dec 8-16	- Request review and extraction of final results
Dec 17	- Project report

From the workplan provided above, the early steps were actually followed based on the planned time. The project itself was approved within a week, including the IRB determination process. One notable step that was modified was the data preprocessing, in terms of cleaning vendor names. While initially planned, the time was instead spent on figuring out how to generate the edgelist from SQL queries. Additionally, my supervisor also indicated that the vendor names used in the database had already gone through significant cleaning, but will still have duplicates due to encoding issues: in that each university submitting to IRIS will most likely have a different way of encoding a vendor or corporation.

Additionally, one of the significant issues encountered during the planning was concerns on data access. Because the data was stored in a secure data enclave, I could only work on the project during office hours, only learning much, much later that using TeamViewer allowed for remote work. Because of this, most of the work done for this project was compressed to the 1st week of December. Additionally, the security of the data meant that tools for community detection, such as the Python-Louvain package, or the Stanford Network Analysis package in Python, could not be used in the enclave, necessitating around 1 week in coordination to extract the necessary data to work on the project locally.

Acknowledgments

I would like to kindly acknowledge the help of my direct supervisor, Ms. Natsuko Nicholls of IRIS, who guided me in the whole approval process of the project, and in extracting the data for community detection analysis.

References

J. Folkstad and S. C. Hayne. 2011. Visualization and analysis of social networks of research funding. In *2011 44th Hawaii International Conference on System Sciences(HICSS)*. volume 00, pages 1–10. https://doi.org/10.1109/HICSS.2011.487.

Hakan Kardes, Abdullah Sevincer, Mehmet Hadi Gunes, and Murat Yuksel. 2013. Complex network analysis of research funding : A case study of nsf grants.

Julia Lane, Jason Owen-Smith, Rebecca Rosen, and Bruce Weinberg. 2014. New linked data on research investments: Scientific workforce, productivity, and public value. Working Paper 20683, National Bureau of Economic Research. https://doi.org/10.3386/w20683.

News Service. 2017. U-m annual research expenditures reach new high. https://news.umich.edu/u-m-annual-research-expenditures-reach-new-hig

A Supplemental Material

Community	Nodes	Ave. degree	Freeman	Ave. clustering
Connected Component	44774	56.4339	0.1693	0.7499
Louvain Communities				
Community 8	7735	32.7007	0.2767	0.7378
Community 4	7560	38.6920	0.2894	0.7214
Community 1	5870	36.9952	0.4447	0.7478
Community 0	5564	30.4597	0.5531	0.7431
Community 3	4336	14.3261	0.3893	0.6791
Community 6	2822	135.5010	0.4741	0.7180
Community 7	1481	12.8872	0.6043	0.7309
Community 5	1218	21.9310	0.3433	0.7646
Community 15	1045	17.2268	0.1149	0.7538
Community 9	1015	9.8123	0.3658	0.7500
CNM Communities				
Community 0	1347	26.5122	0.2139	0.7369
Community 1	11907	33.599	0.1729	0.7036
Community 5	12179	94.5812	0.3907	0.7581
Community 6	11720	30.3568	0.2124	0.7128
Community 7	1359	12.5710	0.2378	0.7694
Community 8	1274	18.7708	0.0970	0.7594

Table 2: Results of baseline and current network

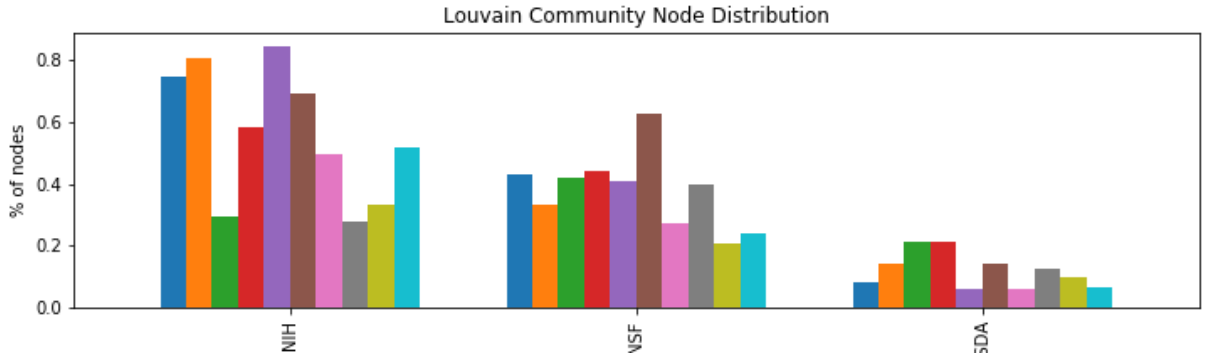

Figure 5: Larger version of Figure 1: Louvain distribution of vendors across NIH, NSF, and USDA

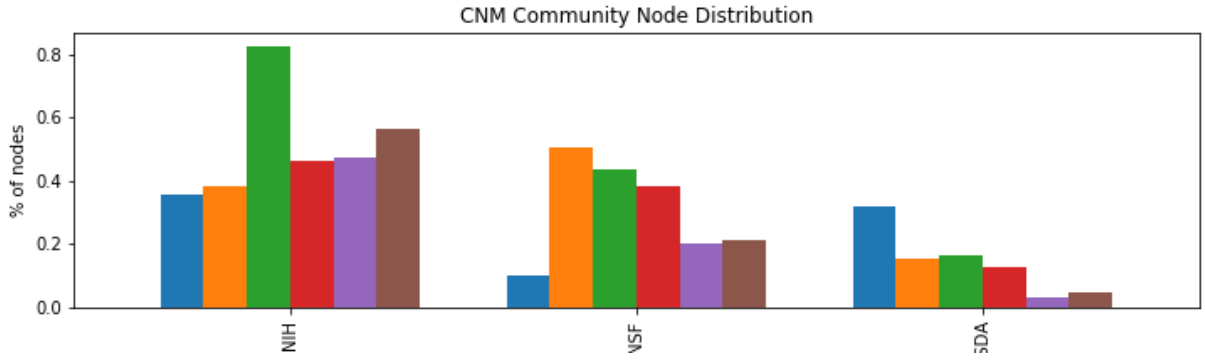

Figure 6: Larger version of Figure 2: CNM distribution of vendors across NIH, NSF, and USDA

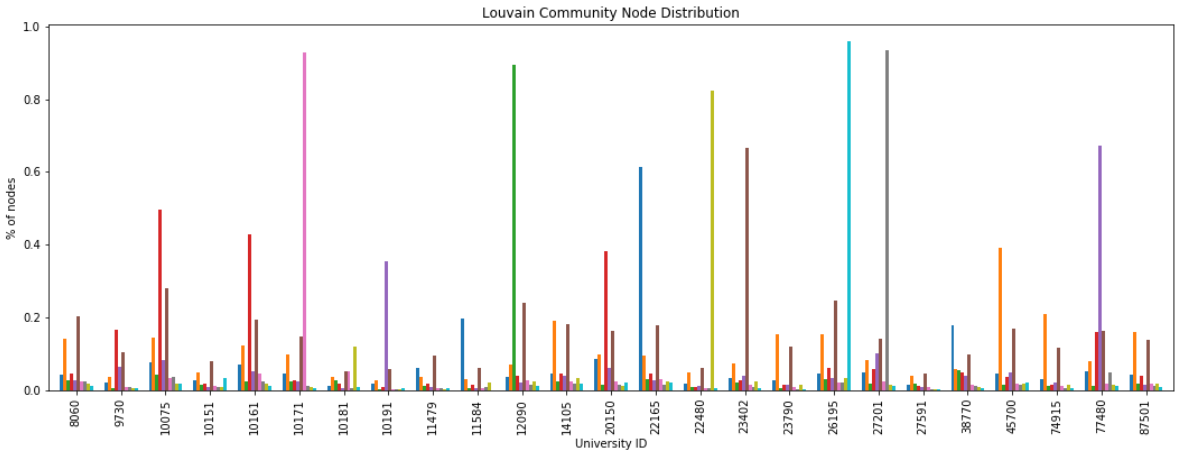

Figure 7: Larger version of Figure 1: Louvain distribution of vendors across universities

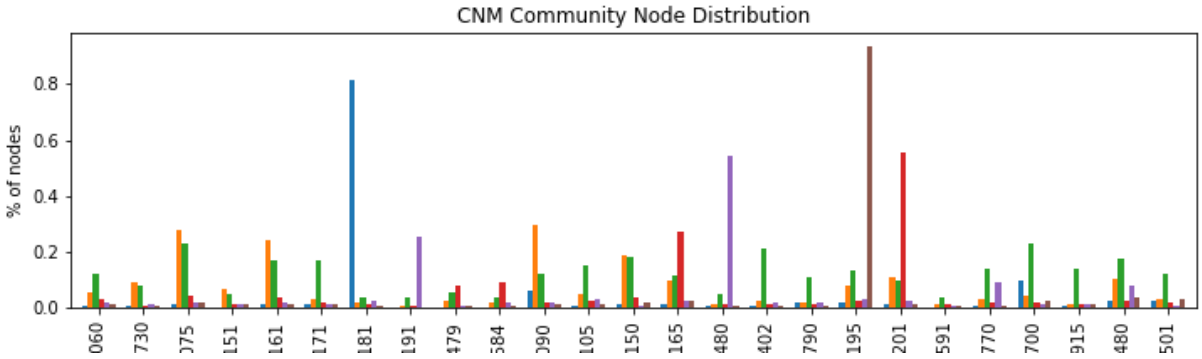

Figure 8: Larger version of Figure 2: CNM distribution of vendors across universities