COMMUNITY DETECTION VS. GROUND TRUTH IN PHYSICIAN REFERRAL NETWORKS

SAJID ZAIDI*

1 INTRODUCTION

Physicians often refer patients to another physician for specialized treatment that they themselves cannot provide. For example, a family practice physician may refer a patient to a surgeon. These patient referrals define a network in which physicians are nodes and the edges denote patient referrals. In this paper, I will use a community detection algorithm to identify clusters of physicians. These clusters might reflect formal entities such as hospitals or physician group practices, or they may reflect more informal associations of physicians in a particular geographic location. After identifying communities using an algorithm, I will then test these communities against the "ground truth" data to see how well the algorithm has identified the true clusters in the data. Community detection can have many applications in the healthcare field, such as identifying the spread of new medical technologies through informal networks. However, before such applications can proceed, community detection algorithms need to be demonstrated to perform well on physician data. This paper will address this problem.

This paper will have two major components: detection of communities from physician network data, and testing of these communities against ground truth data.

2 PRIOR WORK

Relatively few papers have applied the tools of network analysis to healthcare, and even fewer have used administrative claims data. Barnett et al.[1] used shared patient relationships to infer physician networks. A shared patient relationship means that two physicians each saw the same patient at some point during the year. If two physicians share a patient, an edge is created between them, and the paper found that these edges are a good predictor (0.73 area under ROC curve) of whether two physicians actually have a referral relationship (verified using survey). However, one weakness

* *sajidsrzaidi@gmail.com*

of this approach is that sharing patients is a noisy indicator of a relationship between physicians. For example, if a patient sees their cardiologist and then sees a gynecologist 6 months later, these two events are unlikely to indicate any relationship between the two physicians, but they would count as an edge under this method. By using data that directly observes referrals, I am able to avoid this issue.

In a separate paper, Barnett et al.[2] investigate the effect of network structure on hospital spending. Their paper uses claims data to construct a network of physicians using the patient sharing metric described above. They find that the higher the average degree of physicians in a hospital, the higher is the spending in that hospital. They also find that in hospitals in which primary care physicians have more "relative centrality" compared to specialist physicians, there is lower spending and fewer medical specialist visits. One possible flaw with this study is that the outcome measure (spending) is mechanically correlated with the metric of average degree, since patients who see more doctors and therefore cost more money are also more likely to create edges in the network. Despite this issue, the paper illustrates the value of analyzing physician referral networks using network analysis techniques and shows some important applications to health policy issues.

In this paper I implement community detection, and there is a large literature describing methods of generating communities. One popular method for large networks is the Clauset-Newman-Moore (CNM) algorithm [3]. This algorithm optimizes modularity, and uses efficient data structures to improve runtime over older algorithms such as Girvan-Newman [4]. Clauset et al. tested their method on multiple real world graphs with known community structure and showed that their method detects this structure with high sensitivity and reliability. Many other researchers have used the method successfully. For example, Leskovec et al[5] show that the method performs well and achieves qualitatively similar results as the spectral method. Moreover, they showed that CNM is computationally cheaper than many other algorithms and thus scales better to large networks like the physician network I will use. For these reasons, I will use the CNM algorithm, and I describe it in more detail in the Algorithms section.

Finally, there is a small literature on how to compare generated communities against ground truth data. In the past, most community detection papers have been unable to compare performance against ground truth communities because for many networks the data doesn't exist, or otherwise they have used very small datasets such as Zachary's karate club or the NCAA football conference and visually compare generated communities against the ground truth. Recently, some papers have begun to use more quantative methods of comparison. Hric et al. [6] propose two new methods of quantitative performance evaluation and find that community detection methods do not do well in uncovering the ground truth. Yang and Leskovec found similarly poor results [7]. I will describe their algorithms later in the paper, and use their results as a baseline for comparison. However, the poor performance of community detection when applied to a wide range of datasets by both Hric et al. and Yang and Leskovec leads me to expect similarly poor performance in this application.

3 DATA PREPROCESSING

It takes a considerable number of steps to create the network data from the raw claims data. I use a large dataset of administrative health insurance claims from the Medicare program. Medicare is a government health insurance program for people of age 65 and older, as well as disabled persons of any age. This dataset consists of 30 million patients in the fee for service Medicare program, which is a large portion of the population over 65. Each observation in the dataset is a claim that represents a health care service, and contains identification information on the physician who performed the service as well as the physician who referred the patient (if any). Since this dataset is extremely large, I restrict to patients with prostate cancer since it is a disease that mainly affects the elderly, it has a well-defined course of treatment, and it is easy to identify the physicians who treat the disease. I define a cohort of patients by using the Institute of Medicine's algorithm for identifying newly diagnosed cases of prostate cancer [8]. The algorithm identifies patients with a diagnosis in the study year who have no prior diagnosis of prostate cancer in a lookback window of 6 months. From all the claims for this cohort of patients, I restrict to claims specifically treating prostate cancer. Each of these claims lists the physician performing a health care service as well as the physician who referred the patient, so each of these claims defines an edge between physicians.

Although a referral is directed, I will treat these edges as undirected because most referrals are from a primary care physician to a specialist, not the other way around, and thus a directed network would result in the undesirable property that most specialist physicians would have very low out degree. One could also consider these edges to be weighted, with the number of referrals constituting the weights, but I will consider them unweighted for ease of analysis as well as to avoid overweighting physicians who happen to see more patients and thus refer more patients. So, finally, I have constructed an undirected, unweighted graph where the nodes represent physicians and an edge represents at least one patient referral between the physicians. I have data on the employing organization of each of the physicians in the sample, and these employment relationships will be the ground truth communities. This reflects the intuition that physicians in large group practices or hospitals generally refer patients to other physicians in their organization.

Following the methodology of Hric [6], I subset the network to the largest connected component to facilitate analysis. This does not drop very many nodes, as the largest connected component is 94% of the original network. Finally, I also restrict to nodes that are members of a ground truth community with at least 3 members, to avoid including singleton private practice physicians in the data. Table 1 gives descriptive summary statistics of the network I have constructed.

4 ALGORITHMS AND METHODS

I first describe the CNM algorithm [3]. The algorithm greedily optimizes modularity. The algorithm proceeds as follows (from the CNM paper):

1. Begin with each node in its own community

2. Combine the two communities that result in the biggest increase in modularity
3. Recalculate the modularity
4. Repeat from step 2 until there is one overall community

This is a greedy agglomerative method of combining the graph into communities. The partition that gives the highest modularity is returned as the result of the algorithm. I do not describe the calculation of modularity in detail since it was covered in class.

After obtaining the optimal community partition from this algorithm, I will evaluate it against the ground truth employment data using multiple measures. The first measure I will use is Normalized Mutual Information (NMI), which was also used by Hric et al. and was first proposed by Danon et al. [9]. The following description closely follows Danon's paper. For a confusion matrix N, where the rows correspond to ground truth communities and the columns correspond to detected communities, the element N_{ij} is the number of nodes in ground truth community i that appear in detected community j. The NMI is

$$I(A,B) = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} N_{ij}\log(\frac{N_{ij}N}{N_{i.}N_{.j}})}{\sum_{i=1}^{C_A} N_{i.}\log(\frac{N_{i.}}{N}) + \sum_{j=1}^{C_B} N_{.j}\log(\frac{N_{.j}}{N})} \tag{1}$$

where the number of real communities is C_A, the number of detected communities is C_B, the sum over row i is denoted $N_{i.}$ and the sum over column j is denoted $N_{.j}$. This measure gives the amount of information extracted by the algorithm. If the algorithm matches the real communities perfectly, $I(A,B) = 1$ and if the algorithm's results are totally independent of the real communities, $I(A,B) = 0$.

The second measure I will use is average maximal precision and average maximal recall. The following description closely follows Hric et al [6]. Yang and Leskovec[7] use a similar method. Let C_i represent the set of nodes of ground truth community i, and D_j represent the set of nodes of the detected community j. The Jaccard similarity between these two communities is defined as

$$J(C_i, D_j) = \frac{|C_i \cap D_j|}{|C_i \cup D_j|} \tag{2}$$

For each ground truth community, we want to define a recall score. We do this by searching over all the detected communities, and use the most similar detected community to compute the maximal recall. In other words, the maximal recall for ground truth community C_i is

$$R(C_i) = \max_{D_j \in \{D\}} J(C_i, D_j) \tag{3}$$

Similarly, the maximal precision for detected community D_j is

$$P(D_j) = \max_{C_i \in \{C\}} J(C_i, D_j) \tag{4}$$

Since these measures associate a precision and recall score with each community, I will plot these scores in ascending rank order to get a sense of the accuracy of our algorithm. This will be similar to an ROC curve. I will also compute the average maximal precision and average maximal recall, since these are proportional to the area under the curve.

5 RESULTS

The following table provides summary statistics of the network. The number of edges is comparatively low for a network of this size, and this may reflect the fact that any given physician tends to refer patients to a small set of trusted doctors. The clustering coefficient is in line with other real world networks we have examined in class, so it is clear that this network can't be described as a random graph.

Table 1: Network Descriptive Statistics

Nodes	**Edges**	**Ground Truth Communities**	**Clustering Coefficient**
80,890	198,785	9,831	0.1199

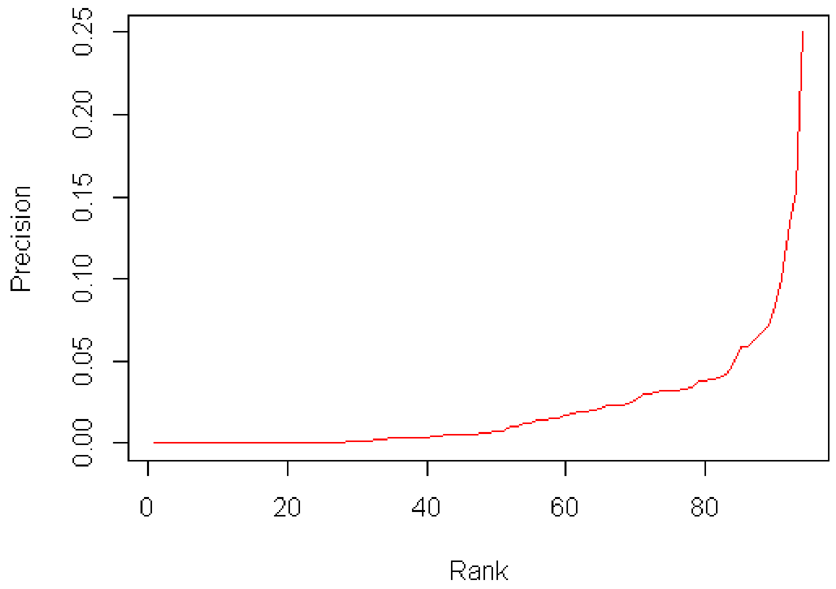
The following table gives the results of the analysis, and plots of maximal average precision and maximal average recall are on the next page.

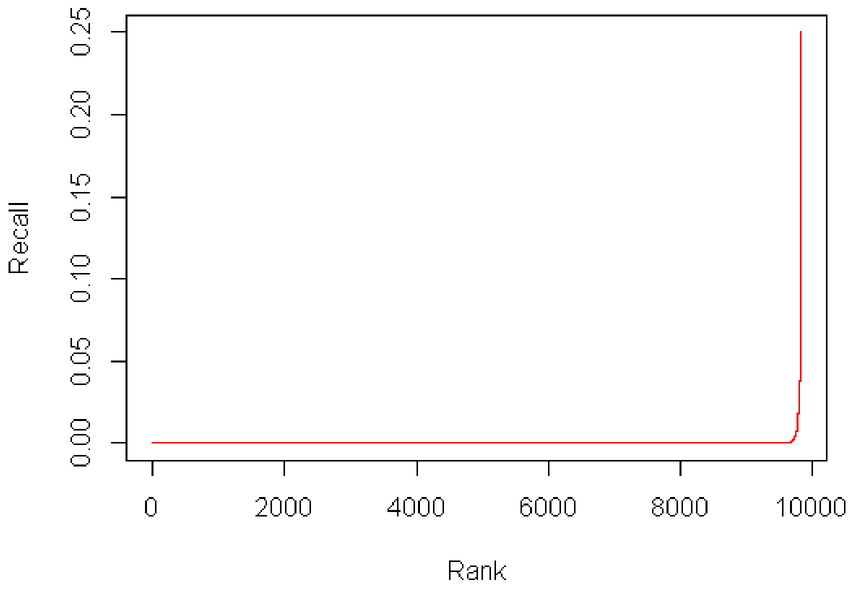
Table 2: Results of Clauset-Newman-Moore

Predicted Communities	**Modularity**	**Normalized Mutual Information**	**Avg. Maximal Recall**	**Avg. Maximal Precision**
94	0.907	0.0538	0.00027	0.021

The first thing one notices is that the number of predicted communities is far smaller than the number of ground truth communities. This may be explained by the fact that the ground truth communities are highly overlapping– most physicians are members in more than one network. The CNM algorithm gives a non-overlapping partition, and it is not surprising that it only takes a small number of non-overlapping communities to cover the highly overlapping ground truth communities.

The performance of the algorithm is very low, as predicted by the literature. Indeed, Hric et al. tested multiple community detection algorithms of different kinds on 10 large datasets, and found that NMI scores rarely go above 0.3, and sometimes are less than 0.1. This is in line with my results, as I find a Normalized Mutual Information of 0.05. The average recall and precision is also very low. The Hric paper found that, using the Louvain algorithm (which is based on modularity and is most similar to CNM) the mean average recall across the 10 large networks they tested was 0.323 and the mean average precision was 0.26. Thus, my results on these two measures are considerably lower than those found in the literature. The ROC curves on the next page also demonstrate the poor performance. The recall curve has almost zero area under the curve, as might be expected since there were over 9,000 ground truth communities but only 94 detected communities, so one would not expected the detected communities to closely match most of the ground truth communities. The precision curve is slightly better, but still demonstrates poor performance.





One possible reason for the poor performance is that my definition of ground truth (employment organization) may not correspond well with my definition of network edges (patient referrals). In future work, I would want to investigate whether most patient referrals are actually to physicians in outside organizations. Another possible reason is that communities among physicians may be constituted in other ways, such as communication, requests for medical advice, collaboration at conferences, medical school alumni networks, etc., and formal patient referrals do not capture these more ubiquitous informal relationships.

6 CONCLUSION

Identifying community structure among physicians can shed light on many interesting questions in health care research. For example, there has been concern in the health policy community that urologists who are part of a large group practice have an incentive to refer patients to their specialist colleagues for expensive treatments because they can share in the profits[10][11]. Another interesting application is to identify influential physicians who spread the use of new medical technologies in their informal networks. These applications and many more depend on accurate identification of communities of physicians, and this paper is the first I am aware of to apply community detection to physician networks and test its accuracy against ground truth data. The ability to accurately detect communities is important because in most cases, ground truth data is not available.

It is possible that an effective community detection algorithm will require the use of non-topological features to supplement the purely topological approach of most current algorithms. In the case of physician networks, using features such as treatment patterns, specialty, location, and many others would likely improve our ability to identify networks of physicians. In any case, more research is necessary to improve our ability to detect physician networks.

REFERENCES

[1] Michael Barnett, Bruce Landon, A. James O'Malley, Nancy Keating, and Nicholas Christakis. Mapping physician networks with self-reported and administrative data. *Health Services Research*, 46:1592–1609, 2011.

[2] Michael Barnett, Nicholas Christakis, James O'Malley, Jukka-Pekka Onnela, Nancy Keating, and Bruce Landon. Physician patient-sharing networks and the cost and intensity of care in us hospitals. *Medical Care*, 50:152–160, 2012.

[3] Aaron Clauset, M.E.J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.

[4] Michelle Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.*, 99:8271–8276, 2002.

[5] Jure Leskovec, Kevin Lang, and Michael. Mahoney. Empirical comparison of algorithms for network community detection. *International World Wide Web Conference*, 2010.

[6] Darko Hric, Richard Darst, and Santo Fortunato. Community detection in networks: structural clusters versus ground truth. *arXiv:1406.0146v1*, 2014.

[7] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. *Proceedings of 2012 IEEE International Conference on Data Mining*, 2012.

[8] Thomas MaCurdy et al. Geographic variation in spending, utilization and quality: Medicare and medicaid beneficiaries. *Institute of Medicine*, 2013.

[9] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics:Theory and Experiment*, 2005:P09008, 2005.

[10] Stephanie. Saul. Profit and questions on prostate cancer therapy. *New York Times*, 12/01/ 2006.

[11] John Carreyrou and Janet Adamy. How medicare 'self-referral' thrives on loophole. *Wall Street Journal*, 10/22/ 2014.