Exploring communities in CMS[1] Open Payments data using Community-Affiliation Graph Model (AGM)

CS224W FInal Report

Anne Parker, Bryan Lewandowski, Chaudary Zeeshan Arif

Abstract: We explore model-based community detection using the Community-Affiliation Graph (AGM) Model on a new ground truth network based on the inaugural 2014 Open Payments Public Use File: Creating Public Transparency into Industry – Physician Financial Relationships dataset. Communities are defined by healthcare company financial relationships with healthcare providers. A financial relationship graph is a projection of the weighted directed bipartite affiliation graph of healthcare company payments to healthcare providers. We consider and model two variations of the financial relationship graph based on two weighting scenarios using payments data. We assess our results by comparing model based community membership and overlap with actual community membership and overlap.

1. Introduction

The topic of community detection has been a strong focus of attention in network analysis research in recent years. As a topic of study, community detection is complex because there has been no unified approach to operationalizing the definition; rather communities are defined and evaluated using a variety of measures and algorithms appropriate to the domain being studied. A further complication arises when communities overlap.

A recent development in community detection research is the use of ground truth information present in the data itself as a means to evaluate the performance of a community detection methods [6]. Leveraging this work [7] developed a Community-Affiliation Graph Model (AGM) for detecting communities and explicitly including community overlap structure. AGM is a flexible generalizable model which has been shown to accurately identify non-overlapping as well as overlapping community structure. We are interested in whether AGM can be an effective model for detecting communities in real world networks of moderate to large size, a next step suggested in [7].

We are adding to this line of research by using AGM to identify communities in a new ground truth network based on the inaugural 2014 Open Payments Public Use File: Creating Public Transparency into Industry – Physician Financial Relationships data. We are interested in community size and overlap structure as indicators of healthcare companies "reach" into the healthcare provider population. We are also interested in the relationships between payments, community size and overlap structure.

The data set also includes information on the dollar value of company-provided payments. This additional data will allow us to study the AGM performance for the projection of weighted affiliation networks where the dollar values are treated as weights.

We will first describe the empirical properties of our Physician Financial Relationships network. Next we will explore AGM as a modeling tool on our healthcare provider financial network under two scenarios. Finally we explore several different weighting scenarios through a series of experiments.

[1] Centers for Medicare & Medicaid Service. http://www.cms.gov/openpayments/

2. Community Structure

We define ground truth communities for the Physician Financial Relationships network to be individual healthcare companies. Healthcare companies provide payments to healthcare providers for many reasons including: having well-recognized experts evaluate their products, building a network of healthcare providers that recommend and/or influence others to use existing products as well as provide a conduit to introduce new products. Thus healthcare providers serve a functional role for healthcare company communities. Healthcare company payment programs are of concern because of the potential for conflict of interests, either real or perceived.

We expect community overlap and we hypothesize that such overlap may be sizeable for some healthcare companies. For example, competition among healthcare companies for influential providers as well as a focus on expanding market share will likely result in overlap.

3.0 Related Research

The AGM model has been shown to accurately model community structure and overlap through a rich set of experiments using AGM-model-generated synthetic data as a baseline to compare against real ground truth networks [5][7]. The focus of these experiments has been to validate the AGM model against sub-graphs of real ground truth networks. We are exploring the potential for AGM as a general modeling tool for detecting community structure and overlap on an entire medium to large network. We will follow the approach used in [5] [7] to set up our ground truth communities and associated provider-provider financial network.

Projection of Affiliation Graph with Weighted Edges: The AGM model [7] was developed using un-weighted bipartite graphs and associated projections. Our data has a natural bipartite structure with directed edges representing payments from healthcare companies to healthcare providers. We have weights available for our initial bipartite graph based on the dollar value of payments.

Our dataset includes natural weights based on the dollar values of healthcare company payments to healthcare providers. The methods used to create projections of weighted affiliation networks has been extensively reviewed by [1]. We integrate our weights through the projection of the weighted affiliation graph and use a slightly modified approach from [2] described as unconditional threshold as follows:

Let $B = (V, C, M, W) = \{(u, c, w_{uc}) \in M, W : u \in V,\ c \in C,\ w_{uc} \in W\}$ where *V* is a set of nodes, *C* is a set of communities, *M* is a set of edges with weights *W* where the weights depend on *u* and *c*. Following [1] we consider two weighting scenarios for each healthcare company/healthcare provider link: (1) *wd* = dollars paid and (2) *wp* = proportion of total company payments.

We generate a graph *G*(*V,E*) from the projection of *B* such that for nodes $u,\ v \in V \quad (u,\ v) \in E$ where *u, v* share a community *c* under the following weight scenarios:

1. Create an edge (*u,v*) if both wd_u , wd_v $\geq$ *threshold*
2. Create an edge (*u,v*) if both wp_u , wp_v $\geq$ *threshold*

Our experiments will focus on exploring different thresholds and their effect on community structure, overlap, and AGM model performance.

Evaluation Metrics: We will use a standard set of model evaluation metrics including *recall*, *precision* and F_1 adapted to accommodate community overlap which occurs when a healthcare provider receives payments from multiple healthcare companies. *Recall* is the ratio of the number of correctly classified communities to the number of communities in ground truth, *precision* the ratio of the number of correctly classified communities to the number of model-identified communities, and F_1 the harmonic mean of *precision* and *recall*.

We adapted our evaluation metrics from [4] who provide a detailed analysis of twenty four evaluation metrics for classification tasks. Following the approach in [4] for multi-labeled communities, let n = the number of nodes and k = the number of health care companies. For a node u, let $Y_u \epsilon \{0, 1\}^k$ where the *j*th entry in Y_u = 1 if node u is a member of community c_j and 0 otherwise. Let $Z_u \epsilon \{0, 1\}^k$ where the *j*th entry in Y_i = 1 if the AGM model predicts node u is a member of community c_j and 0 otherwise. Then,

Recall = $\frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Y_i|}$ *Precision* = $\frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Z_i|}$ F_1 = $\frac{1}{n}\sum_{i=1}^{n}\frac{2|Y_i \cap Z_i|}{|Y_i|+|Z_i|}$

Finally we will use the accuracy in the number of communities metric from [6] as 1 - $\frac{k-pred(k)}{k}$ which is a measure of coverage. Computational details for these metrics are described in detail in section 5.0 -- Model fit and evaluation

4.0 Dataset -- Financial Relationship Network

We use the inaugural 2014 Open Payments Public Use File: Creating Public Transparency into Industry – Physician Financial Relationships data set which contains financial transactions between August 1 through December 2013. This data set is administered by the Centers for Medicare and Medicaid Services (CMS) and made available as a result of the Affordable Care Act. The data include a rich set of financial data including dollar amounts of consulting fees, research grants, travel reimbursements, and other in-kind gifts as well as limited medical provider and medical company demographics [3].

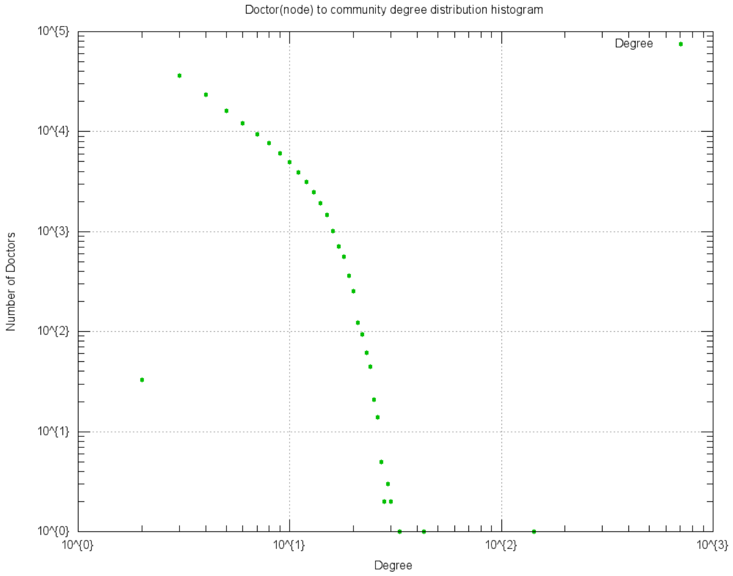
The data set consists of 358,924 healthcare providers and 1,607 healthcare companies. There are approximately 2.6 million payment records however there are often multiple payment records per healthcare provider. When we aggregate multiple payments for each healthcare provider we have approximately 1.1 million payment records spanning 1,135 healthcare companies (ground truth communities). Due to resource limitations we imposed two constraints on our benchmark graph -- we (1) eliminated healthcare companies who made payments to fewer than 3 healthcare providers and (2) eliminated healthcare providers with less than 3 payments from healthcare companies. This reduced the number of healthcare providers by 227,423 and the number of healthcare companies by 176.

Data Preprocessing: We used Pentaho Data Integration engine to ETL data from source, filtered the extra information and transformed into edge list format for consumption by the AGM community detection. The number of transactions were rolled-up (the payment amounts were summed) between a particular provider and company. From this data file, another file was generated to project into a provider-provider network by forming edges between providers if they were paid by the same company. This provider-provider edge list was then fed to the AGM community detector for a final output.

The following table provides some statistics about our filtered Open Payments data set, which we used to generate our bipartite (provider to company) graph.

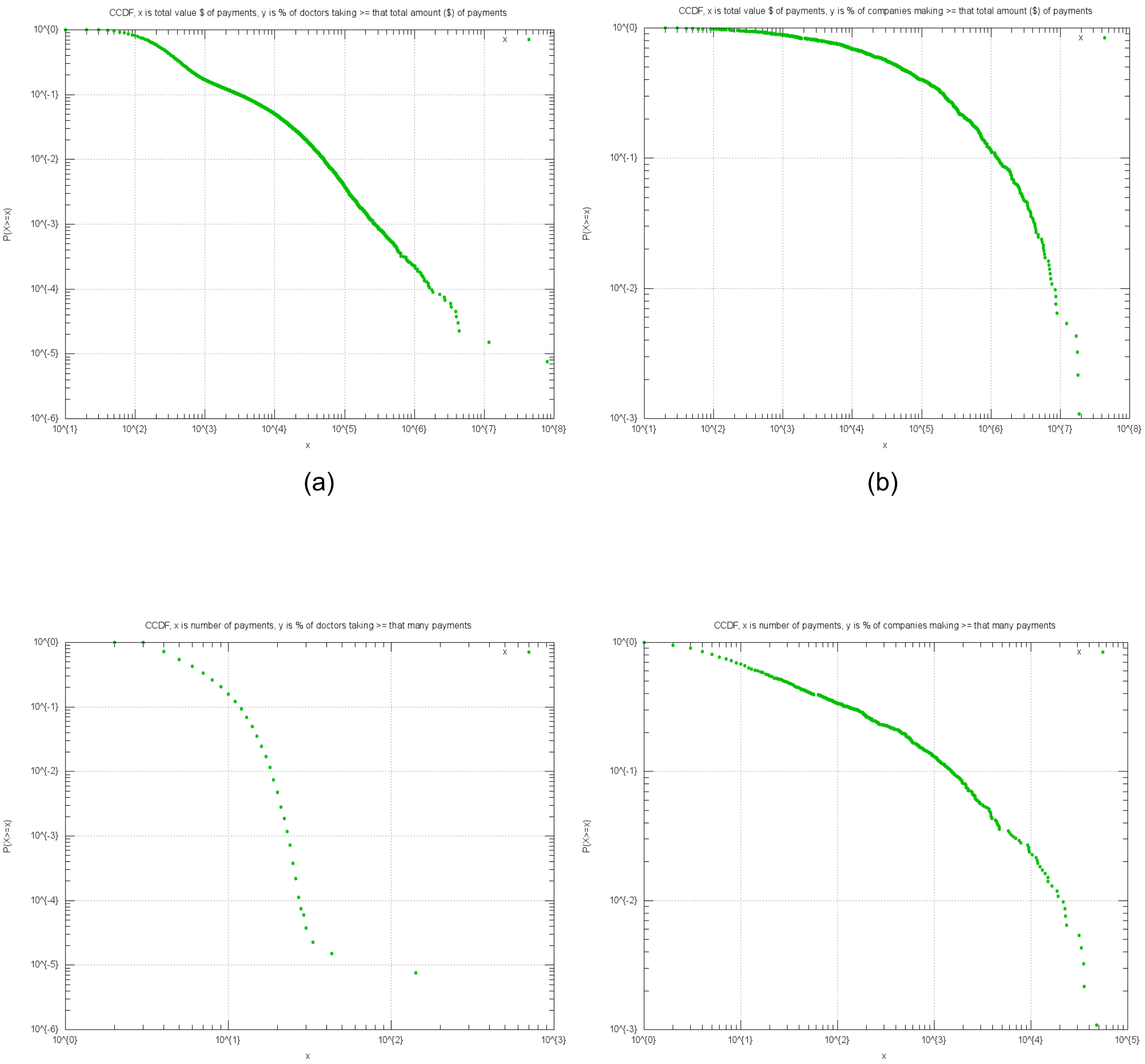
Table 1: Open Payments dataset (filtered) overview

Number of nodes (providers)	18,940	Number of unique provider-company edges	32,283
Number of truth communities (companies)	752	Average community size	43
Average number of distinct companies a provider receives from	1.7	Total of all payments	$464,552,389.11



The complementary cumulative distribution function (CCDF) of the ratio of providers taking an amount of money is shown in Figure 1(a), and the CCDF of the total amount of money spent by companies is shown in Figure 1(b). CCDF plots in terms of number of payments (from different companies, or to different providers) are shown in Figure 1(c) and 1(d).

Figure 1: CCDFs of the payment amount and number of payment sources.



(a)

(b)

(c)

(d)

Figures 1(b) and 1(d) show that our dataset has an exponential falloff in terms of both the number of providers that a company pays and the total dollar amount paid by companies to all providers. Figure 1(a) shows that the amount of money taken by providers is similar to a power law, however figure 1(c) shows an interesting distribution of community membership which is neither exponential nor a power law. This makes the dataset interesting as it exhibits both exponential and power law behavior.

5.0 Model Fit and Evaluation

We attempted to fit our full affiliated bipartite projection with un-weighted edges using the AGM model. The agmfit and BigClam portions of the SNAP network analysis toolset [2] were used. However due to computing resource limitations and dataset size we were unable to run the AGM model on our baseline provider-provider network. Our solution was to apply a dollar threshold of $1,000 which reduced the graph to a size such that we were able to successfully run the model. Table 2 provides an overview. We verified that imposing this constraint preserved the baseline characteristics of the network.

Table 2: Reduced baseline dataset overview

Number of nodes (providers)	18,940
Number of truth communities (companies)	752
Number of unique provider-company edges	5,827,276

We ran AGM under two community detection scenarios -- automatic AGM model community detection and forced AGM model community detection with the number of communities pre-specified as the number of ground truth communities. A complication in the computation of evaluation metrics is the need to connect model-generated communities to ground truth communities (healthcare companies). We used the following procedure:

1. Constructed a HashMap of <Company, [List of MDs paid by that company]>.
2. AGM model results identify communities and the providers who make up the communities. We capture this data as List<List of providers>. The outer list contains the size of each AGM community and the inner list enumerates the providers in each community.
3. We relate the company with the detected community by iterating over the list from step 2 and loop through the HashMap from step 1. When a match is found, the list from step 1 is added to a new list of all possible providers that could be part of the matched community.

The collections obtained from step 3, for e.g. an intersection of step 2 and List from step 3 gives us all the providers that are detected successfully. The remaining ones are possibly missed. Such collections and their size ratios give us the evaluation metrics. The evaluation metrics for each scenario are provided in Table 3.

Table 3: Evaluation metrics for AGM model detection scenarios

	Number of communities detected	Community size range	Precision	Recall	F1	Accuracy in the number of communities
Automatic AGM	100	6 - 1969	0.9970	0.0269	0.0525	0.1324
Forced AGM (detect 748 communities)	738	3 - 1859	0.9983	0.0195	0.0383	0.9853

Precision is high indicating that AGM detected communities are identified accurately, however the low recall indicates AGM fails to detect many communities and those that are detected, while accurate, may not be complete. Although our forced AGM model identified 737 communities, 334 of those communities were randomly filled by the algorithm.

6.0 Experiments

Our experiments focused on exploring different weighting thresholds and their effect on community structure, overlap, and AGM model performance. Our weighting scenarios used different sets of thresholds to identify influential providers and communities and are defined as follows: Let *C* be equal to the percent of a company's payments and *D* be equal to the percent of a provider's payments. With different threshold values for *C* and *D* we measure the congruence of significant payments (as measured by the payment representing *C*% or more of the company's total payments to all providers) and a significant payment to the provider (as measured by such company payments representing *D*% or more of the total payments received by that provider from all companies).

We ran six such experiments finding both high precision and recall but low accuracy in the number of communities. Our results are provided in Table 4.

One limiting factor is a dramatic drop-off in the average degree of a provider in the provider-company bipartite network as the values of C and D are increased. This value represents the overlap in communities; with a value of 1 there is no overlap at all (no multi-community providers). The value of D sets an effective upper bound on the degree of a provider in the bipartite network of 1 / D. Setting the value of C much higher than 1 cuts off vast amounts of the network. This is illustrated in Table 4, compare the entries to the value in the unfiltered dataset of 1.7. We attempted experiments to run BigClam and our procedure on the full dataset, but the network was simply too large for BigClam / AGM to process. Future work to develop AGM algorithms with improved performance would provide a good opportunity to revisit our work.

We noted that the ground truth bipartite graph generated from our weighting procedure was different from our baseline network in that the average number of distinct companies a provider receives payments from is close to 1, significantly lower than the baseline data set. This suggests that there are likely many disconnected single components as well as a few larger components. This would account for the high precision. We are uncertain how AGM handles disconnected components, if at all. We noted that in [5][7] disconnected components from the same group are treated as separate ground truth communities. However we will never have the foregoing condition in our graph by construction since any provider paid by the same company will always be connected by an edge.

Table 4: Evaluation metrics for AGM model detection experiments using weighting thresholds

	Average number of truth communities a provider belongs to	Number of communities detected	Community size range	Precision	Recall	F1	Accuracy in the number of communities	Number of ground truth communities after weighting
C = 1% *D* = 25%	1.0312	100	6 - 74	1.00	0.9969	0.9984	0.0961	1,041
C = 1% *D* = 33%	1.0182	100	6 - 74	1.00	0.9954	0.9977	0.0828	1,208
C = 10% *D* = 25%	1.0181	71	3 - 6	1.00	0.9542	0.9766	0.0904	785
C = 10% *D* = 33%	1.0091	69	3 - 6	1.00	0.9886	0.9943	0.0910	758
C = 25% *D* = 25%	1.0107	8	3 - 4	1.00	1.00	1.00	0.0150	534
C = 25% *D* = 33%	1.0064	7	3 - 4	1.00	1.00	1.00	0.0138	508

7.0 Summary and Conclusions

We explored AGM as a modeling tool for identifying communities in a set of FInancial Relationship Networks derived from the 2014 Open Payments Public Use File: Creating Public Transparency into Industry – Physician Financial Relationships data. Overall our experiments had very high precision but very low recall suggesting that the model accurately identifies some communities and overlap but may fail to identify a majority of communities.

Our assessment of AGM as a general modeling tool for identifying community structure and overlap based on our experiments with a single real world network is mixed. We know that AGM accurately detects community structure including overlap, however it lacks coverage. However, we don't know whether the detected communities capture the "essence" of our problem objective. For our specific network we are interested in community size and overlap structure as indicators of healthcare companies "reach" into the healthcare provider population. Further we are interested in identifying the most influential provider(s) identified by healthcare companies. These influential provider(s) correspond to the notion a connector node in [5][7]. For our baseline provider-provider network we have one connector node with a degree of 5,854. This provider had payments from 17 companies totaling $76,898.79 as per ground truth. AGM also identified this provider in the overlap of 17 detected communities.

Precision is an important metric for our network because wrongly assigning a provider to a community is perhaps a more critical error than failing to include all providers of a particular community (as defined by ground truth). While coverage is low are the communities detected by AGM the largest communities? Subsequent work is needed to address these questions and is outside the scope of this paper. Future work with better infrastructure to support large storage (>500GBs of edgelist data) and clustered setup to take advantage of BigClam (an AGM implementation for parallel execution) will result into detection of communities on complete dataset i.e. without any thresholds into improving the computational performance of the AGM model would enable this work to be revisited with a richer data set.

8.0 References

[1] Neal, Zachary. "The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors." *Social Networks* 39 (2014): 84-97.

[2] Leskovec, Jure and Sosi, Rok. SNAP: A general purpose network analysis and graph mining library. http://snap.stanford.edu/snap, June 2014.

[3] OpenPayments Public Use Files: Methodology Overview & Data Dictionary, The Centers for Medicare & Medicaid, September 2014

[4] Sokolova, Marina, and Guy Lapalme. "A systematic analysis of performance measures for classification tasks." *Information Processing & Management* 45.4 (2009): 427-437.

[5] Yang, Jaewon, and Jure Leskovec. "Community-affiliation graph model for overlapping network community detection." *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 2012.

[6] Yang, Jaewon, and Jure Leskovec. "Defining and evaluating network communities based on ground-truth." *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*. ACM, 2012.

[7] Yang, Jaewon, and Jure Leskovec. "Structure and overlaps of communities in networks." *arXiv preprint arXiv:1205.6228* (2012).