Exploration of Link Structure and Community-based Node Roles in Network Analysis

Jerry Scripps, Pang-Ning Tan, and Abdol-Hossein Esfahanian
Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
{scripps,ptan,esfahanian}@cse.msu.edu

Abstract

Communities are nodes in a network that are grouped together based on a common set of properties. While the communities and link structures are often thought to be in alignment, it may not be the case when the communities are defined using other external criterion. In this paper we provide a new way to measure the alignment. We also provide a new metric that can be used to estimate the number of communities to which a node is attached. This metric, along with degree, is used to assign a community-based role to nodes. We demonstrate the usefulness of the community-based node roles by applying them to the influence maximization problem.

1 Introduction

Networks are deceptively simple data representation, with related objects connected by directed or undirected links. Yet they may give rise to a surprising number of interesting problems—from determining the influence or authoritativeness of objects to finding communities in large networks. While the field of network analysis has been extensively studied for many years, new problems and solutions continue to emerge from the analysis of networked data.

In this paper, we argue that the utilization of community knowledge can lead to a valuable extension of existing network analysis methods. We first explore the connection between the link structure and the communities present in networked data. Many existing community finding algorithms implicitly assume that the communities are compatible with the link structure of a network. Intuitively, this is supported by our familiarity with social networks where groups of people who are interconnected by friendship can be considered to be a community. We consider the communities that are aligned with the link structure as *natural communities* of the network. In practice, however, a network can have communities imposed by other factors beyond its link structure. For example, with social networking sites, members can join special interest groups independent of the friends they have established. If the communities and link structure are incompatible, it would be meaningless to apply link mining on the network to learn characteristics of the non-natural communities.

Newman and Girvan [2] proposed a metric called modularity function (Q) to compute the fit between the community and link structure. Although the metric seems quite intuitive, it has several limitations which make it inappropriate for comparing the compatibility of communities in two networks. To overcome these limitations, we propose a pair of statistics for measuring the alignment between communities and link structure. A new metric called rawComm is also introduced to estimate the number of communities to which a node is attached. This metric, along with degree, is used to define community-based node roles. Roles, which determine the function that nodes play with respect to the network, can convey meanings such as authority, popularity, and influence. Our node role definition is the first that we are aware of that makes use of community knowledge. Finally, we illustrate how community-based roles can extend current methods of influence maximization [1].

2 Community and Link Structure Alignment

A community refers to nodes that are grouped together based on a common set of properties. The alignment between the community and link structure of a network indicates how well nodes within the same community are linked and those in different communities are not linked. The modularity function Q computes this fit by comparing the actual number of intra-community links to its expected value when the network is randomly wired. Values of Q close to one indicate strong alignment between the link and community structure whereas values close to zero indicate no alignment. The measure however has several inherent limitations. First, the range for Q is sensitive to the number of communities and number of nodes in each community. The more communities a network has the narrower the range of possible values for Q will be. Thus, Q is an adequate measure when comparing algorithms on the same network with the same number of communities but not for comparing one network to another or comparing different number of communities for the same network. Q also does not explicitly take non-links into consideration.

Our proposed alignment metrics are defined based on the concepts of pure and complete node pairs [3]. A node pair is complete if it is linked and belongs to the same community. A node pair is pure if it is not linked and belongs to different communities. We define a pair of alignment metrics: p, which measures the fraction of linked node pairs that are complete and q, which measures the fraction of non-linked node pairs that are pure. For example, a p value of 0.9 means 90% of the links involve nodes from the same community. On the other hand, a q value of 0.9 means only 10% of the node pairs belonging to different communities are linked.

There are a number of differences between p and q and the modularity function Q. First, the meaning of p and q are clear—p is the fraction of links within communities and q is the fraction of non-links between communities. High values of p and q indicate that the link structure is well-aligned with the communities. Also, their ranges are well defined; both can take values between 0 and 1. This makes p and q more appropriate statistics for comparing different algorithms, different networks or different community sets for a single network than Q. Additionally, unlike the modularity metric Q, the q statistic explicitly takes non-links into consideration. For a more thorough comparison of the metrics, readers may refer to our technical report [4].

3 Community-Based Node Roles

In social network analysis, roles are used to describe the behavior of a node in relationship to its neighbors and to the network at large. Examples of node roles include those that are based on their popularity (degree), centrality (closeness and betweenness) and authority (e.g., PageRank).

3.1 Community-Based Role Definition Community-based roles are useful in a number of ways. First, they provide useful information to analysts in areas as such as anti-terrorism and law enforcement. In searching for potential terrorist threats, for example, analysts may find it useful to identify suspects with certain roles (mastermind, financier, facilitators, military commander, etc). If they were looking for persons with diverse contacts, they could focus on nodes whose community-based roles are designated as bridges or ambassadors (see Figure 1). Second, community-based roles could also be utilized in existing link mining applications such as ranking, link prediction, and influence maximization, and node classification. An example illustrating the application of community-based roles to influence maximization is given in Section 4.

We define the community-based role of a node according to the number of communities and links incident to it. Figure 1 shows a community-degree chart that is divided into four quadrants for the four different roles. The vertical axis represents the degree while the horizontal axis represents the community metric. The community-based node role is identified based on which of the four quadrants a node falls into. Nodes in the upper right quadrant are those with a high degree and a high community score. They act as *ambassadors*, providing connections to many different communities. The upper left quadrant contains what we call *big fish* from the cliche ”big fish in a small pond” meaning that they are very important only within a community. This is due to their having a high degree but a relatively small community score. In the lower right quadrant are those with a low degree but a high community score. These we call *bridges* because they serve as bridges between a small number of communities. Finally, in the lower left are the *loners*—those with a low relative degree and low community score.

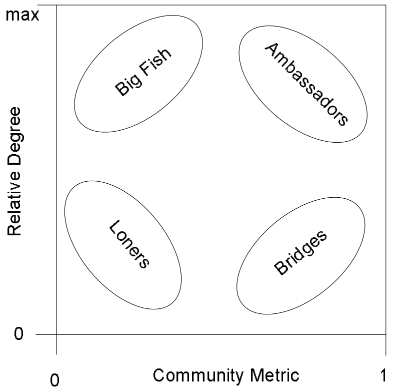

Figure 1. Community-based roles

The metrics shown in the diagram have been normalized to values between 0 and 1. For the community metric, we subtracted the minimum and divided by the range between maximum and minimum. For degree, we divided by the highest degree node in the network, giving us a relative degree score between 0 and 1. In our experiments, we chose a threshold of .5 to classify the node roles; however, depending on the distributions of degree and community metric scores, other thresholds can be chosen.

In order to define the community-based node role it is necessary to measure the number of communities linked to each node. If the community membership information is available, this can easily be done. However, often it is not available, in which case a method is needed to estimate it from the network. The next section presents our proposed community metric known as rawComm.

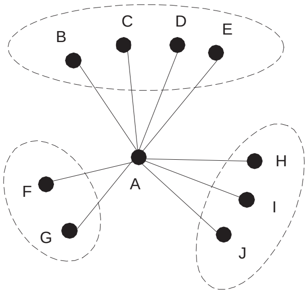

Figure 2. Community membership contribution of neighbors to rawComm of node A.

3.2 rawComm Metric To understand the intuition for our metric, consider the diagram shown in Figure 2. There are 9 nodes from three communities in the neighborhood of node A. Nodes B through E are attached to one community, F and G are attached to another, and H through J are attached to a third community. Another way of computing the number of communities connected to A is to add up the *community membership contributions* from each of its neighboring nodes. For example, since B is in a community of four nodes (excluding A), its community membership contribution to A is $\frac{1}{4}$. Similarly, F is in a community with one other node, so it is assigned a community membership contribution of $\frac{1}{2}$, and so on. When all of the community membership contributions from nodes B to J are added, the total is 3, which is exactly the number of communities to which node A is attached.

We now formalize our method for computing the community metric, which we call rawComm. The community metric for the node i, is defined as follows:

$$\text{rawComm}(i) = \sum_{j \in N(i)} \tau_i(j) \tag{3.1}$$

where $N(i)$ is the set of nodes in the neighborhood of node i, and $\tau_i(j)$ is the community membership contribution of node j to the rawComm of node i. Following the above discussion, we define the community membership contribution $\tau_i(j)$ as follows:

$$\tau_i(j) = \frac{1}{1 + C_{ij}} \tag{3.2}$$

where C_{ij} is the number of nodes (other than j) that are neighbors of i and are in the same community with j. For example, since B is in a community with three other nodes, $C_{AB} = 3$ and $\tau_A(B) = 1/4$.

Note that Equations 3.1 and 3.2 give the exact formula for computing the community metric for node i even though it is computed using only nodes that are adjacent to i. If we know the community membership of every node, then it is possible to compute the community metric precisely. Otherwise, we need to estimate the parameters $\tau_i(j)$ and C_{ij} from the topology of the network. A probabilistic approach for estimating the expected values of these parameters are given in the next section. We show that the approximations become reliable when the communities are well-aligned with the network topology.

3.3 Estimating Community Membership Contribution Let $\mathcal{G} = (V, E)$ be a graph and $v \in V$ is one of its nodes. The neighborhood of v, denoted as $N(v) = (V_v, E_v)$, is an induced subgraph of $\mathcal{G}$, where $V_v = \{u \in V | (u, v) \in E\}$ and $E_v = \{(u, w) | u \in V_v, w \in V_v, (u, w) \in E\}$. Our method for estimating the community membership contribution τ is based on the following two assumptions. First, we assume that the community structure and network topology are well aligned (i.e., p and q are at least larger than 0.5). The second assumption is that the neighborhood of a node contains all the information necessary to estimate the community membership contribution. Following these assumptions, the next two theorems give the formula for computing the expected value of $\tau_v(u)$ using the neighborhood information $N(v)$.

THEOREM 3.1. *Given a node v and its neighborhood $N(v) = (V_v, E_v)$, the expected value of the community membership contribution for node $u \in V_v$ is*

$$E[\tau_v(u)] = \sum_{g \in \Omega_v} \frac{P(g)}{|g|} \delta(u \in g) \tag{3.3}$$

where $P(g)$ is the probability of a community g, $|g|$ is the community size, Ω_v is the set of all possible community assignments, and δ is an indicator function whose value is 1 if its argument is true and 0 otherwise.

Theorem 3.3 states that the expected value of the community membership contribution of node u can be computed by the sum of probabilities of all communities containing node u weighted by the community size. A proof for this theorem can be found in our technical report [4]. To compute the right hand side of Equation 3.3, we need to enumerate every community g that contains u and compute $P(g)$, which is a very expensive procedure. We propose a more efficient approach by building a probability model that uses the links between u and the other nodes in V_v.

THEOREM 3.2. *Given a node v and its neighborhood, $N(v) = (V_v, E_v)$, the estimated contribution of node u to v's community count is*

$$E[\tau_v(u)] = \frac{1}{1 + n_1 p + n_2(1 - q)} \tag{3.4}$$

where n_1 is the number of nodes in V_v linked to u and n_2 is the number of nodes in V_v not linked to u.

The proof for this theorem can be found in our technical report. In order to apply Equation 3.4 we need to know the values of p and q for a given network. Unless otherwise mentioned, our experiments were conducted using $p = q = 1$. As we will show in Section 5.2, even if the approximation is not very good we can still achieve good results for two reasons. First, we are mainly interested in communities that are consistent with the network links, which means that their p and q values should be reasonably high. Otherwise, we should not expect the metric, nor any community finding algorithms, to produce a view that agrees with the community concept we have in mind. Second, to define the type of community-based node roles, it is sufficient to know the relative ordering of their community scores rather than their absolute magnitude. As long as the metric does better than random guessing, we expect to see some improvements in the experimental results.

4 Application of Roles to Influence Maximization

Influence maximization is concerned with finding the most influential nodes in a network. We assume that the nodes in the network are capable of adopting an idea, purchasing a product or something similar. This process is referred to as activating. We also assume that nodes that are activated may influence their immediate neighbors who themselves may choose to activate. The problem becomes choosing the best nodes to initially activate in order to maximize the number of activated nodes at the end of the process. Kempe et al. introduced several models to describe the behavior of the node activation [1]. Our experiments use the Independent Cascade model, in which influence is spread from node to node in discrete time steps. A node i that becomes active in step t has one chance to make its inactive neighbors active in step $t + 1$. The probability that node i will activate node j is given by the edge weight.

Current work in this area is focused on maximizing the raw number of nodes activated. We propose extending the problem to focus on the number of communities covered. A community is covered if one of the nodes in the community is activated. Our approach is to choose the initial set of nodes using the community-based node roles in order to maximize the number of communities covered. The first method selects those nodes with the highest rawComm score which focuses on ambassadors and bridges. The second method focuses exclusively on ambassadors by choosing nodes with a combination of high rawComm and high degree. The results of our experiments show that using roles to maximize community coverage shows improvement over the other influence maximization methods.

5 Supporting Analysis and Experiments

This section presents the empirical evidence that demonstrates the usefulness of our proposed approaches.

5.1 Data Sets Our experiments were conducted on three data sets—movie data from UCI KDD repository, Enron email data, and the Facebook social networking data. For the movie data set, we created a link between actors who have co-starred in at least one movie together. The resulting network had 3,725 nodes and 58,123 links. The Enron data set is a collection of email messages exchanged among 149 executives at Enron. We formed a network by establishing links between executives who had more than five email exchanges between them. The FaceBook data set was created from crawling the FaceBook web site for a medium sized university in Michigan. FaceBook is a social networking site that allows students to join, post pictures, text and other descriptive information. Most importantly they can create links to friends and join groups. The resulting network contains 2,550 nodes.

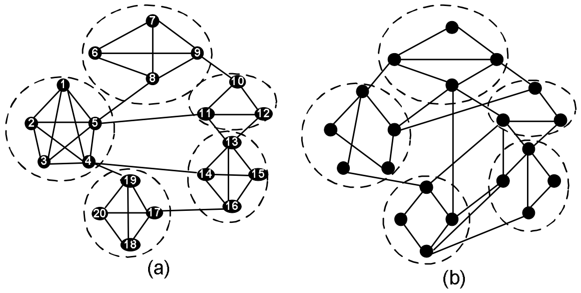

Figure 3. Effect of p on τ

In addition to the three real data sets we also ran our experiments on the synthetic networks shown in Figure 3. Both networks contain the same node set and communities but different link structures. We purposely fixed the communities so that one could clearly see the difference in the alignment between the link structure in (a) versus (b).

5.2 Community and link structure alignment The goals of our experiments in this section are to show that:

- p and q provide an effective measure of the extent to which communities are aligned with the link structure
- p and q often provide more meaningful information about the alignment than modularity (Q)
- rawComm provides a reliable estimate for community metric when p and q are sufficiently high.

For each of the three real data sets, we calculated the p, q and Q values for two types of community definitions: (1) based on some externally imposed criterion and (2) based on the results obtained using the normalized cut (nCut) spectral clustering algorithm [5].

For the FaceBook data we create three types of communities, each containing 20 communities. The first com-

Table 1. Alignment of Communities with Link Structure

Grouping method	Nbr of comm.	p	q	Q	SSE	Avg error	Avg comm.
FaceBook							
age, etc.	20	0.12	0.91	-.19	53k	3.70	10.95
hometown	20	0.58	0.46	-.31	36k	2.89	7.47
major	20	0.59	0.49	-.25	32k	2.67	7.06
nCut	20	0.44	0.93	.23	17k	2.01	9.15
Enron executives							
title	12	0.16	0.87	-.28	1124	2.25	4.55
nCut	12	0.61	0.94	.36	256	0.99	3.07
Movie actors							
genre	14	1.00	0.16	1.11	183k	6.64	11.01
nCut	14	0.75	0.70	.23	35k	2.13	2.73
Synthetic							
net (a)	5	.79	1.00	0.48	1.31	0.25	1.75
net (b)	5	0.59	0.93	0.19	11.78	0.53	2.20

munity type groups the students by a combination of gender, age and political party preference. Next, the students are grouped by their hometown (19 most popular and other) and after that, by their concentration (or major). For Enron, the employees were grouped by title (director, vice-president, etc). For the movie data set, actors were placed in genres based on the movies they appeared.

5.2.1 Comparison between Community Alignment Measures

To see how p and q can be used to characterize the alignment of link and community structure, consider the results shown in Table 1. We will concentrate on the p, q and Q columns of this table. Observe that communities defined using nCut are often more compatible with the link structure than communities defined using external criterion. This is not surprising since nCut uses the link structure to create its communities.

For FaceBook, observe that the community formed using a combination of age, gender, and political leaning has a very low value of p (almost 90% of the links are between communities) but a high value of q (more than 90% of non-links are between communities). This means that two friends are almost certain to be in different communities based on age/gender/politics but also that two non-friends are also very likely to be in different communities. We conclude that these communities are not aligned well with the links but the non-links are aligned. For the next two community definitions—hometown and major—the alignment is better for p but for q it is about 50/50. All three externally defined community types have very poor alignment but the p and q values are very different for age/gender/politics compared to hometown and major. Yet, the value of Q is about the same for all three. Thus, even if the alignment is poor, p and q allows us to distinguish between data sets where Q does not. As another example, note that the nCut communities for both FaceBook and movie had a Q of 0.23. This would imply that the alignment between the communities and the link structure was of a similar nature but as we can see by looking the the p and q values this is not the case. As mentioned in Section 2, the magnitude of Q may not be meaningful since it is sensitive to the number of communities and number of nodes in each community. Instead it is more suitable as a relative measure for comparing different alignments on the same network with the same number of communities.

The Enron data has a high q and low p for communities defined based on the executive's title. From this we can surmise that executives are very likely to email other executives outside of their community but also that executives who do not email each other are most likely from different communities. For movies, the links between actors appear to be good predictors for communities defined according to movie genre. This is partly because the communities are overlapping (in our technical report [4] we explain how p and q are modified to handle overlap). Since the q value is low, this implies that even if two actors did not appear together in a movie they will still likely work in the same genre.

For the synthetic data, we observe that both networks have fairly high values for both p and q but that (a) appears to be better aligned than (b). Visually inspecting the networks confirms this. The Q value also supports this but does not let us know why. We can tell from p and q that it is the links (p) that cause the misalignment and not the non-links (q).

5.2.2 Effect of alignment on rawComm

The last three columns of Table 1 correspond to the following: (1) **SSE**, which is the sum-squared error between the number of actual communities linked to each node and the number predicted by rawComm, (2) **Avg error**, which represents the average of the absolute value of the difference between the actual number of communities and rawComm of all nodes, and (3) **Avg Comm**, which is the average number of actual communities assigned to the nodes by each of the community definition methods.

For each data set, we compare the average SSE found by each grouping method. The smaller the average SSE is, the more accurate our estimation of rawComm is. For FaceBook, nCut has the best SSE which is not surprising given that it has a high q value and an average p. Looking at the rest of the table it is clear that with better alignment (higher p and q values), the accuracy of rawComm improves.

The last two columns reveal another view of the accuracy of rawComm. For FaceBook, using nCut, rawComm, on average, is off by about two where the average is about nine communities per node. This seems quite a reasonable estimate, considering that the p value is less than 0.5.

Table 2. Comparisons of Community Influence Maximization techniques

	Movies			Enron		FaceBook		
algorithm	**nodes**	**Director**	**Genre**	**nodes**	**group cov.**	**nodes**	**group cov.**	**time (ms.)**
random	11	60	9.7	13	5.3	12	3.5	10
degree	19	83	12	13	6.9	21	3.7	60
greedy	22	259	12	15	6.8	23	4.1	4,920,000
comm	18	261	12	14	8.4	18	5.6	371
ambass	20	288	13	13	7.5	20	4.3	411

5.3 Influence Maximization The influence maximization problem is implemented using the independent cascade model, with the edge weights set to .01. Due to the stochastic nature of the model, for each method, we repeated the simulation 5,000 times and then averaged the results. For evaluation purposes, we compared the total number of nodes that are activated as well as the number of communities reached using five different algorithms. The baseline *random* approach selects k nodes randomly. *degree* selects the k nodes with highest degree. The algorithm proposed by Kempe, et al. [1], labeled *greedy*, chooses one new node each iteration, selecting the node that will result in the greatest increase of activated nodes according to the Independent Cascade model. The last two methods use the metrics proposed in this paper. The method *comm* selects the k nodes with the highest rawComm score and *ambass* selects the k nodes with the highest sum of normalized degree and rawComm. The nodes selected by *ambass* are mostly ambassadors while those selected by *comm* is a combination of ambassadors and bridges.

The results for the movie data are included in Table 2. The first column labeled *nodes* is the average number of nodes activated by the initial 10 nodes. The columns under *Director* and *Genre* indicate the number of communities that had at least one node activated. The greedy method, not surprisingly, was able to activate the greatest number of nodes. However, even though *ambass* activated fewer nodes, it was able to reach more communities. *comm* also was able to spread to a large number of communities even though it selected fewer nodes than *degree*, *greedy* or *ambass*. That is not too surprising given that nodes connected to many communities are not necessarily high degree nodes. The p values were 1 for both types of community definitions. The q values were .947 and .164 for director and genre communities respectively. This means that for both types of communities if there is a link between two actors there is a 100% chance that the two actors will be in the same community. If there is no link between them then the q tells us that, for the director set, it is almost certain that they will not be in the same community, whereas for genre, there is only a 16% chance that they will be in different communities.

For Enron data, because of the sparsity of the network, we used an edge weight of .05. Again, even though *comm* and *ambass* did not activate the most nodes they provided the widest coverage. For the FaceBook data we used communities based on concentration (students major). Once again the algorithms *comm* and *ambass* outperform the others on spreading to a larger number of communities. For this data set we also kept track of the execution time for the different algorithms which is listed in the last column. It can be seen that all the algorithms are quite fast with the exception of the *greedy* algorithm with is approximately 12,000 times slower than the slowest of the others. With large data sets or when time is critical, *degree* or *ambass* would be worthy alternatives.

6 Conclusions

In this paper we explore the advantages of utilizing community knowledge in the analysis of networked data. Towards this end, we introduce several community-based roles according to the number of communities and the degree of each node. We show that nodes with roles called ambassadors are useful to maximize the number of communities activated during influence maximization. Metrics for estimating the number of communities and measuring the compatibility between communities and link structure are also proposed in this study.

References

[1] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. *Conference on Knowledge Discovery in Data*, pages 137–146, 2003.

[2] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69.

[3] J. Scripps and P. N. Tan. Clustering in the presence of bridge-nodes. *Proc of SDM'06: SIAM Int'l Conf on Data Mining, Bethesda, MD*, 2006.

[4] J. Scripps, P.-N. Tan, and A.H. Esfahanian. Exploration of link structure and community-based node roles in network analysis. Technical Report TR, Michigan State University, 2007.

[5] J. Shi and J. Malik. Normalized cuts and image segmentation. *Ieee Transactions On Pattern Analysis And Machine Intelligence*, 22(8), August 2000.