Community Detection in a Large Real-World Social Network

Karsten Steinhaeuser[1] and Nitesh V. Chawla[2]

[1]ksteinha@cse.nd.edu, University of Notre Dame, IN, USA
[2]nchawla@cse.nd.edu, University of Notre Dame, IN, USA

Abstract Identifying meaningful community structure in social networks is a hard problem, and extreme network size or sparseness of the network compound the difficulty of the task. With a proliferation of real-world network datasets there has been an increasing demand for algorithms that work effectively and efficiently. Existing methods are limited by their computational requirements and rely heavily on the network topology, which fails in scale-free networks. Yet, in addition to the network connectivity, many datasets also include attributes of individual nodes, but current methods are unable to incorporate this data. Cognizant of these requirements we propose a simple approach that stirs away from complex algorithms, focusing instead on the edge weights; more specifically, we leverage the node attributes to compute better weights. Our experimental results on a real-world social network show that a simple thresholding method with edge weights based on node attributes is sufficient to identify a very strong community structure.

1 Introduction

Modern data mining is often confronted with the problems arising from complex relationships in data. In social computing, the analysis of social networks has emerged as an area of great interest. On one hand, such interaction networks offer an advantage because they can represent rich, complex information in an intuitive fashion. On the other hand, mining this information can be quite difficult as many existing methods are not directly applicable to network data, and graph theoretic algorithms are computationally very expensive. Therefore, there is an immediate need for efficient algorithms to analyze social networks.

We address one particular task in social network mining, namely *community detection*. A number of methods to address this problem have been proposed, and Newman distinguishes these into two categories: bottom-up "sociological" approaches and top-down "computer science" approaches; a more detailed treatment with examples of each is provided in [5]. Both have been shown to perform well in practice, but regardless of the fundamental approach most algorithms are computationally expensive. Their scalability is limited to at most a few thousand nodes as execution becomes intractable for larger networks [7]. However, datasets containing millions of nodes are becoming readily available, and analyzing them requires highly scalable algorithms.

In this work we take advantage of the important social tendency of *homophily* – "birds of a feather flock together" – to analyze a cellular phone network, which is a unique real-world social network in that it consists of 1.3 million individuals connected by actual communication patterns between them. The search for community structure is guided by a similarity function based on attributes attached to nodes in the network, not just the topology. We believe the latter is limiting as it does not carry the important element of "closeness" among neighbors. Our hypothesis is that using node attributes (in this case demographic information about the individuals) to compute edge weights is sufficient to identify communities in the network, whereas weights computed by other means produce significantly inferior results. We show that a relatively simpler and highly scalable algorithm is able to produce extremely high modularity scores, surpassing empirical limits specified by Newman [6].

The remainder of this paper is organized as follows: In section 2, we describe three different similarity metrics used to weight the edges of a network. In section 3 we present the setup and experimental evaluation on a real-world social network. Finally, in section 4 we conclude with a discussion of the results and their implication for social network analysis.

2 Edge Weighting Methods

We assume that the connectivity of the network is provided as part of the input. If this is all the information given, then the only criterion the algorithm can consider is the network topology, i.e. measurements like clustering coefficients, shortest paths, etc. Yet it is often the case that rich information about nodes and/or edges is available, which allow us to assign them more meaningful weights. In this section, we describe three different methods for weighting the edges of the network: two topological metrics and one based solely on node attributes.

2.1 Terms and Notation

Here we briefly introduce some terms and notation that are used throughout the ensuing discussion. A network is defined as graph $G = (V,E)$ consisting of a set of n nodes V and a set of m edges E between them. Letters i, j, v refer to nodes; $e(i, j)$ denotes an edge connecting nodes i and j, while $w(i, j)$ specifies the weight of the edge. For practical reasons, the graph is represented as an adjacency list such that $neighbors(i)$, the set of all nodes connected to i, is readily accessible. If node attributes for i are available, they are stored as $i.1, i.2, ..., i.r$.

2.2 Clustering Coefficient Similarity (CCS)

Several node similarity metrics are described in [1]. We adopt the topological *clustering coefficient similarity (CCS)* for our work. As the name indicates, the underlying computation requires finding the clustering coefficient *CC* of node v,

$$CC(v) = \frac{2n_v}{d_v(d_v - 1)}$$

where n_v is the number of triangles v participates in and d_v is the degree of v. The weight $w_{ccs}(i, j)$ is computed as a similarity between two nodes i and j, defined as the difference in their clustering coefficients with (CC) and without ($CC^{'}$) the edge connecting them present,

$$w_{ccs}(i, j) = CC(i) + CC(j) - CC^{'}(i) - CC^{'}(j)$$

Intuitively, CCS measures the contribution of $e(i, j)$ to the connectedness among the immediate neighbors of nodes i and j. Algorithm 1 shows the procedure for weighting the entire graph using this similarity metric.

Algorithm 1 Clustering Coefficient Similarity (CCS)

```
1: for each node i = 1...n do
2:     w(i, j) = 0
3:     for each node j = 1...neighbors(i) do
4:         w(i, j) = w(i, j) + CC(i)
5:         w(i, j) = w(i, j) + CC(j)
6:         remove e(i, j)
7:         w(i, j) = w(i, j) − CC(i)
8:         w(i, j) = w(i, j) − CC(j)
9:         re-insert e(i, j)
10:    end for
11: end for
```

2.3 Common Neighbor Similarity (CNS)

The second metric is based on a quantity known in set theory as the *Jaccard Coefficient*. For sets P and Q, it is computed as the ratio of the intersection to the union of the two sets. To compute the weight of edge $e(i, j)$, simply substitute $neighbors(i)$ and $neighbors(j)$ for P and Q, respectively, which results in the ratio between the number of neighbors two nodes share (common neighbors) and the total number of nodes they are (collectively) connected to,

$$w_{cns}(i, j) = \frac{|neighbors(i) \cap neighbors(j)|}{|neighbors(i) \cup neighbors(j)|}$$

This metric, called *common neighbor similarity (CNS)*, is intended to capture the overall connectedness among the immediate neighborhood of nodes i and j. Algorithm 2 shows the procedure for weighting the entire graph with the CNS metric.

Algorithm 2 Common Neighbors Similarity (CNS)

1: **for** each node i = 1...n **do**
2: **for** each node j = 1...$neighbors(i)$ **do**
3: $w(i,j) = \frac{|neighbors(i) \cap neighbors(j)|}{|neighbors(i) \cup neighbors(j)|}$
4: **end for**
5: **end for**

2.4 Node Attribute Similarity (NAS)

Note that both of the previous metrics rely solely on the network topology. We postulate that a similarity metric that takes into account node attributes can produce more meaningful weights, thereby improving the community structure. One choice for this scenario might be the Heterogeneous Value Distance Metric [8], but since there is no concept of class among the nodes it cannot be applied directly. However, we can adapt its premise to the situation at hand. We propose to weight edges based on a *node attribute similarity (NAS)* computed as follows: for each nominal attribute a_c, if two connected nodes have the same value then increment the edge weight by one,

$$if\ i.a_c = j.a_c,\ w_{na}(i,j) = w_{na}(i,j) + 1$$

For continuous attributes, to find the weight of edge $e(i,j)$ we first normalize each attribute to (0,1) and then take the arithmetic difference between the pairs of values attribute values to obtain a similarity score. More formally, for each continuous attribute a_n,

$$w_{na}(i,j) += (1 - \alpha|i.a_n - j.a_n|)$$

where α is a normalizing constant. This metric captures the edge weight as the attribute-similarity of two connected nodes. Algorithm 3 shows the procedure for weighting the entire graph using this heterogeneous NAS metric.

Algorithm 3 Node Attribute Similarity (NAS)

```
1: for each node i = 1...n do
2:     for each node j = 1...neighbors(i) do
3:         w(i, j) = 0
4:         for each node attribute a do
5:             if a is nominal and i.a = j.a then
6:                 w(i, j) = w(i, j) + 1
7:             else if a is continuous then
8:                 w(i, j) = w(i, j) + 1 − α|i.a − j.a|
9:             end if
10:        end for
11:    end for
12: end for
```

Finally, note that it is possible to bypass the edge weighting, for instance if weights are known a priori and provided as part of the input. We give an example of this scenario in the following section.

3 Experimental Evaluation

In this section, we provide an example of a large real-world social network and present an experimental evaluation of the weighting methods discussed above. But first we introduce a validation metric, which allows us to quantify the quality of community structure in networks and enables a comparison between the different weightings.

3.1 Validation Metric

When the true structure of a network is known, the Adjusted Rand Index (ARI) is an appropriate validation metric [4]. In real-world networks this is often not the case, and other criteria must be used instead. Newman and Girvan propose a measure rooted in the notion that nodes within a community should be more tightly connected, while nodes in different communities should share relatively fewer connections [6]. The metric, called *modularity*, takes the fraction of within-community edges minus the expected value of the same number of edges placed at random, summed over all communities. For a network with k communities, it is computed using a $k \times k$ matrix where each element d_{ii} denotes the fraction of edges within community i, and d_{ij} the fraction of edges between communities i and j. Modularity is then given by

$$M = \sum_i (d_{ii} - (\sum_j d_{ij})^2)$$

The value normally ranges from 0 to 1 (higher is better) and can vary widely for different real-world networks. Newman et al. report that in social networks it generally falls between 0.3 and 0.7 [6], but there is no threshold value that necessarily seperates "good" from "bad" community structure.

3.2 Community Detection Method

To identify communities in the network, we first apply one of the edge weighting methods to the network and normalize the edge weights to the range (0,1). We then obtain communities using a simple thresholding method. Given threshold t in the same range (0,1), we place any pair of nodes i and j whose edge weight exceeds the threshold, i.e. $w(i, j) > t$, in the same community.

3.3 Cellular Phone Network

We evaluate our hypothesis that edge weights are a critical foundation to good community structure on a real-world social network constructed from cellular phone records [3]. The data was collected by a major non-American service provider from March 16 to April 15, 2007. Representing each customer by a node and placing an edge between pairs of users who interacted via a phone call or a text message, we obtain a graph of 1,341,960 nodes and 1,216,128 edges. Unlike other examples of large social networks, which are often extracted from online networking sites, this network is a better representation of a true social network as the interaction between two individuals entails a stronger notion of *intent* to communicate. Given its large size, the cellular phone network is quite unique in this regard.

As shown in Figure 1, the degree distribution in the network approximately follows a straight line when plotted on a log-log scale, which is indicative of a power law. This is one of the defining characteristics of a *scale-free network* [2]. Community detection in this class of networks is particularly difficult as nodes tend to be strongly connected to one of a few central hub nodes, but very sparsely connected among one another otherwise. Using topological metrics, this generally results either in a large number of small components or a small number of giant components, but no true community structure. We show that weighting based on node attributes can help overcome this challenge.

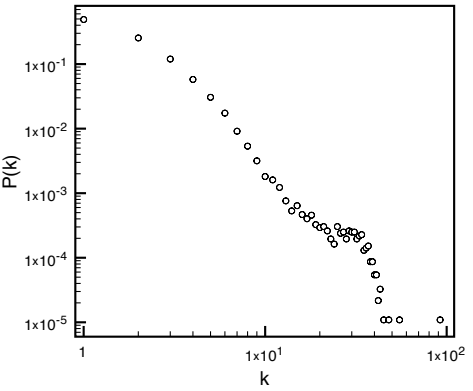

Fig. 1 Degree distribution for the phone network. The presence of a power law indicates that it is a scale-free network.

3.4 Experimental Results

Table 1 shows the effect of threshold t on modularity using the three different edge weighting methods; the execution time for each trial was approximately 40 seconds. We see that this simple thresholding method is sufficient to detect community structure as the modularity values are quite high across the full range of t, although lower thresholds produce better results. In fact, the values over 0.91 for NAS far exceeds the range (0.3,0.7) reported by Newman et al. [6], indicating very strong communities. This shows that the attribute values alone contain some extremely valuable information about the community structure as the NAS metric results in very high modularity.

Table 1 Effects of varying threshold t on modularity with different weighting methods.

Weighting	$t = 0$	0.2	0.4	0.6	0.8
CCS	0.050	0.042	0.020	0.006	0.001
CNS	0.090	0.061	0.022	0.004	0.001
NAS	0.917	0.917	0.917	0.915	0.508

In contrast, the topological weighting methods fail to detect any community structure at all. In this case, the metrics produce many edges with zero weight, fragmenting the network into many thousand singletons and pairs, eliminating the defining characteristics of the community structure.

4 Conclusions

We have explored the viability of various edge weighting methods for the purpose of community detection in very large networks. Specifically, we have shown that edge weights based on the node attribute similarity (i.e. demographic similarity of individuals) are superior to edge weights based on network topology in a large scale-free social network. As witnessed by the fact that a simple thresholding method was sufficient to extract the communities, not only does the NAS metric produce more suitable edge weights, but *all* the information required to detect community structure is contained within those weights. We achieved modularity values exceeding empirical bounds for community structure observed in other (smaller) social networks, confirming that this approach does indeed produce meaningful results. An additional advantage of this method is its simplicity, which makes it scalable to networks of over one million nodes.

References

1. S. Asur, D. Ucar, S. Parthasarathy: An Ensemble Framework for Clustering Protein-Protein Interaction Graphs. In Proceedings of ISMB (2007)
2. A.-L. Barabási and E. Bonabeau: Scale-free networks. Scientific American 288 (2003) 50–59
3. G. Madey, A.-L. Barabási, N. V. Chawla, et al: Enhanced Situational Awareness: Application of DDDAS Concepts to Emergency and Disaster Management. In LNCS 4487 (2007) 1090–1097
4. G. Milligan, M. Cooper: A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis. Multiv. Behav. Res. bf 21 (1986) 441–458
5. M. E. Newman: Detecting community structure in networks. Eur. Phys. J. bf B38 (2004) 321–330
6. M. E. Newman: Finding and evaluating community structure in networks. Phys. Rev. E bf 69 (2004) 023113
7. P. Pons, M. Latapy: Computing communities in large networks using random walks. J. of Graph Alg. and App. bf 10 (2006) 191–218
8. D. R. Wilson, T. R. Martinez: Improved heterogeneous distance functions. J. Art. Int. Res. bf 6 (1997) 1–34