CT 2.3.1

Is Modularity the Answer to Evaluating Community Structure in Networks?

K. Steinhaeuser[1], N. V. Chawla[2]

[1]Dept. of Computer Science & Engineering, University of Notre Dame, 326 Fitzpatrick Hall, Notre Dame, IN 46556, USA
[2]Dept. of Computer Science & Engineering, University of Notre Dame, 384 Fitzpatrick Hall, Notre Dame, IN 46556, USA; also affiliated with the Center for Complex Networks Research, University of Notre Dame, 225 Nieuwland Science Hall, Notre Dame, IN 46556, USA

Introduction

One important task in network analysis is that of *community detection*, or partitioning of a network into groups of nodes that belong together. The problem is relevant to a variety of application areas, but social networks in particular have been the subject of research at the intersection of physics, computer science, and the social sciences. In this context the notion of community is intuitive to grasp, but there is no consensus on a formal definition of the concept. Our work addresses this issue in two primary ways:

(i) We examine *modularity* (Newman, 2004) as both an evaluation measure of community structure as well as an optimization criterion used by some algorithms to identify communities in networks (Clauset, 2004; Duch, 2005). Specifically, we assess the pros and cons of modularity, and identify its shortcomings via comparison to alternate evaluation metrics on networks for which the true communities are known.

(ii) We develop a simple method for community detection using random walks (Figure 1) and show that it can identify the actual communities at least as well as more complex algorithms (Table 1).

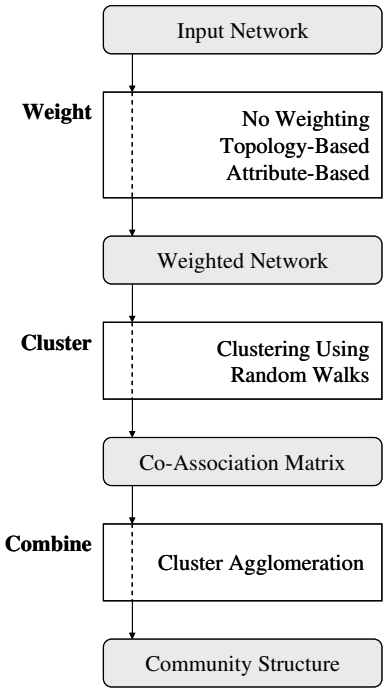

Fig. 1. Algorithm framework for community detection using random walks

We further expand upon the idea of community detection with random walks by extending the method to weighted networks and explain how to incorporate *node attributes* – information that is frequently available but usually ignored by other algorithms – to compute edge weights that can aid in detecting more meaningful communities. To demonstrate the scalability of the random walk approach and examine the effect of using node attributes for edge weighting, we apply the method to a real-world social network constructed from cell phone records consisting of 1.3 million customers.

Table 1. Algorithm Complexity

Algorithm	**Complexity**
Fast Modularity (Clauset, 2004)	$O(n \log^2 n)$
WalkTrap (Pons, 2006)	$O(n^2 \log n)$
MCL (van Dongen, 2004)	$O(n^3)$
Random Walks (this work)	$O(n)$

Results

In our first experiment we compare community detection algorithms on several small networks for which the true structure is known. The results are evaluated using modularity as well as other metrics which measure the degree of agreement with the true communities (Table 2). We make two important observations. First, the true community structure does not necessarily correspond to the highest modularity, which is problematic for algorithms that maximize modularity. Second, the random walk method performs as well as or better than the other methods at identifying the true community structure. In our second experiment we show that due to its low complexity, the algorithm can process a network of over 1 million nodes in 40 seconds. We also compare different edge weighting methods and show that using node attributes to compute the weights can result in significant improvements over other methods.

Table 2. Comparison of Algorithms and Metrics
(best value for each dataset+metric shown in *italics*)

Modularity (Q)			
Algorithm \ Dataset	**Karate**	**Risk**	**Football**
Fast Modularity	*0.381*	*0.625*	0.577
WalkTrap	0.360	0.624	*0.604*
MCL	0.359	0.617	0.596
Random Walks	0.371	0.623	0.598
Accuracy for Assigning Node Labels			
Algorithm \ Dataset	**Karate**	**Risk**	**Football**
Fast Modularity	0.971	0.929	0.591
WalkTrap	0.941	0.929	*0.939*
MCL	0.971	0.929	*0.939*
Random Walks	*1.000*	*0.979*	*0.939*
Adjusted Rand Index (ARI)			
Algorithm \ Dataset	**Karate**	**Risk**	**Football**
Fast Modularity	0.882	0.834	0.492
WalkTrap	0.772	0.832	*0.915*
MCL	0.883	0.815	*0.915*
Random Walks	*1.000*	*0.927*	*0.915*
Normalized Mutual Information (NMI)			
Algorithm \ Dataset	**Karate**	**Risk**	**Football**
Fast Modularity	0.837	0.894	0.732
WalkTrap	0.498	0.848	*0.935*
MCL	0.836	0.834	*0.935*
Random Walks	*1.000*	*0.955*	*0.935*

Discussion

Our experimental results show that the maximum modularity does not necessarily correspond to the true communities in networks. And while modularity may be the only evaluation metric currently available, its use as an optimization criterion for community detection can be problematic as such algorithms may converge on a suboptimal solution. We believe the exploration of alternate methods for identifying and evaluating community structure remains an open area of research. We also address the issue of computational complexity and propose an approach to community detection using random walks that is capable of identifying communities effectively and efficiently.

Acknowledgements

We would like to thank Mark Newman for making several datasets publicly available, and László Barabási for providing the cell phone network. Special thanks to the USMA for sponsoring our attendance at NetSci'08 through a travel grant.

References

A. Clauset, M. Newman, C. Moore (2004): Finding community structure in very large networks. Phys. Rev. E 70, 066111.

J. Duch, A. Arenas (2005): Community detection in complex networks using external optimization. Phys. Rev. E 72, 027104.

M. Newman, M. Girvan (2004): Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113.

P. Pons, M. Latapy (2006). Computing communities in large networks using random walks. J. Graph Alg. App. 10(2), 191-218.

S. van Dongen (2004): Graph Clustering by Flow Simulation. Ph.D. thesis, Univ. of Utrecht.