Investigating the Effectiveness of EVC and PCC as Community-Centered Centrality Measures

Kristofferson Culmer
Department of Computer Science
University of Missouri
119 Engineering Building West
Columbia, MO 65211, USA
krcd58@mail.missouri.edu

Dr. Wenjun Zeng
Department of Computer Science
University of Missouri
119 Engineering Building West
Columbia, MO 65211, USA
zengw@missouri.edu

Katie Allmeroth and Jordan Cowart
Department of Computer Science
Southeast Missouri State University
1 University Plaza
Cape Girardeau, MO 63701, USA
{kmallmeroth1s, jrcowart1s} @semo.edu

Abstract

As online social networks continue to become an ever pervasive part of the human social experience, the value of their study also increases. Of particular interest is the ability to determine important, or central, nodes in a network. Current measures of centrality, such as Eigenvector Centrality (EVC) and Principal Component Centrality (PCC), take into account a nodes "connectedness" but do not explicitly consider the diversity of resources available to it: its social capital. It is our goal to investigate the usefulness of EVC and PCC as a community centrality measure and eventually develop a new centrality measure if needed.

1. Introduction

The ability to distinguish the most important nodes within a network is a widely discussed topic and has resulted in a large number of algorithms to accomplish this task; these are known as centrality measures. With regards to social networking, however, these measures fall short. The current measures focus on analyzing the connectedness of nodes within a graph, with more connections resulting in a higher score.

In a social network, connectedness holds a great deal of relevance in deciding the importance of a node but we propose that it should not be the only factor in forming this decision. Other factors to consider include: the nodes inclusion to important communities, its connections with members of other important communities, and its access to a diversity of resources within the network. We consider the combination of the all these factors as the social capital of a node. The purpose of this paper is to give a more formal and rigid definition of this concept of social capital as well as describe a method by which to score it.

The rest of the paper is organized as follows: Section 2 provides an overview of works related to this study. Section 3 discusses the current centrality measures that are used in network analysis. Section 4 discusses our methods of data collection along with the analysis and visualization of this data. Section 5 proposes a framework for a new centrality measure. Section 6 summarizes our conclusion to our work. Section 7 discusses the future work we hope to achieve.

2. Related Work

2.1 EVC and PCC

[1] discussed the need for and development of PCC. EVC assigns centrality based on the strength of the most dominant feature in the data set. Because of this nature, the highest rated nodes are clustered near each other, allowing for a majority of the network to lack centrality information except for those higher rated nodes. Thus, Ilyas et al. introduced the idea of a new centrality measure, PCC, which is inspired by the KLT and principal component analysis. Instead of only focusing on the most dominate feature, PCC takes into consideration additional features. To demonstrate the shortcomings of EVC and the potential use for PCC,

Ilyas et al. used a small data set and computed centrality information using both tools. Using this data, Ilyas et al. evaluated the advantages and disadvantages of using EVC and PCC.

In [2], Ilyas et al. applied EVC and PCC to two large data sets instead of just one small data set. The first set used was an undirected, unweighted friendship graph from Google's Orkut social networking service. In this data set, there were 70,000 users with almost 3 million edges. Results from running EVC and PCC on this data set concluded that using PCC in conjunction with a node selection criterion ([2] used local maxima) identified many more influential nodes in a network than possible by using just EVC. The second data set was a weighted, undirected gaming graph of matches between users of Facebook's 'Fighters Club' application. It contained 667,560 recorded matches between 143,020 users making for a graph with over 500,000 edges between users. By applying EVC and PCC to the data set, Ilyas et al. was able to demonstrate that the addition of more features in PCC adds new social hubs to the list of previously identified hubs without replacing these hubs.

2.2 Community

Yang et al. [3] discussed the use of ground-truth (user defined) communities as a means of benchmarking a community detection algorithm's effectiveness. The correlation between likes and community membership were discussed in [5]; using this information we will assume that an individual liking a page is a declaration of membership within that particular community.

2.3 Web Crawler

Xiao et al. [4] developed a new algorithm for gathering data from Facebook by use of scrapping instead of using Facebook's API, and demonstrated its effectiveness versus the API method. From their research, it was discovered that Facebook's API was too restrictive for unauthorized developers; however web crawlers were still a great way to gather data from Facebook. There was one problem with using a web crawler: Facebook would only display a certain number of friends at one time. At the time of the paper, only 60 friends were returned from a person's friends list. To bypass this limit, Xiao et al. examined Ajax code embedded in an individual's friend page and discovered a way to use the code to extract the entire friend's list. Along with Ajax manipulation, Xiao et al.'s web crawler features a breadth-first search strategy and was tested successfully.

3. Current Centrality Measures

There are currently many methods of determining a node's measure of centrality [6]; however the two that hold the most interest in determining a node's social capital are Eigenvector Centrality (EVC) and Principal Component Centrality (PCC).

EVC is a popular centrality measure used within the social sciences, with PCC essentially being a tunable extension of EVC [1]. EVC quantifies centrality by recursively analyzing the weight and number of connections between nodes; this method has a notable shortcoming however. Because of the recursive nature of the algorithm and the use of only the largest eigenvector, the results are typically skewed to a particular region of the graph [1]. Since social capital should take into account a node's ability to access and influence to important resources and communities across a network, the scope needs to be expanded, hence the interest in PCC.

PCC broadens the idea of using eigenvectors as a key tool in determining centrality. However, instead of using solely the dominant eigenvector, as EVC does, PCC uses the P most dominant feature vectors (eigenvectors) where P $\leq$ the total number of eigenvectors. This method highlights the P most important friendship communities instead of only using scores from the dominant eigenvector. To illustrate this difference, we analyzed the Arxiv GR-QC collaboration network provided by [7] with EVC and PCC. Highlighted are the top 25 scoring nodes for each feature vector analyzed. As stated, EVC only analyzes the dominant feature vector hence the single cluster in Fig. 1. We chose to analyze the five most dominant feature vectors with PCC and this resulted in the five highlighted clusters in Fig. 2.

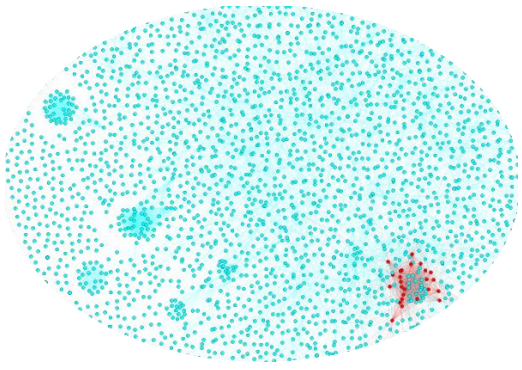


Figure 1 – EVC

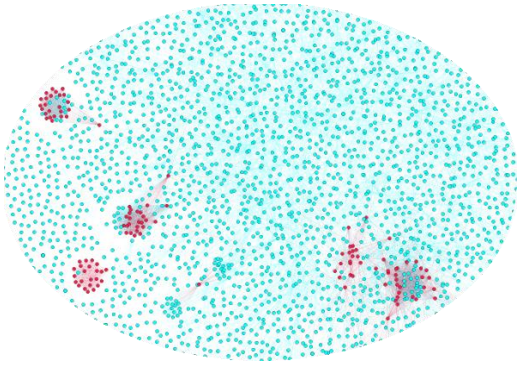


Figure 2 – PCC

4. Data Collection and Analysis

4.1 Data Collection

Data was obtained from Facebook using a crawler based on the algorithm discussed in [4] and was implemented in Python. The crawler was refined in order to only scrape Facebook pages of individuals that attend the University of Missouri (Mizzou). We gathered the following information from these pages: friends that attend Mizzou, gender, academic major, hometown and Mizzou related likes. All user identifiable information was anonymized to preserve privacy. The process for crawling is pictorialized in Fig. 3 and described in Algorithm 1.

The crawler was run on a server running Ubuntu 12.04.4 with a 2.93 GHz Intel Xeon X3470 processor and 4GB of RAM for 72 hours and completely crawled 298 profiles using a targeted crawl. The crawler is capable of crawling approximately 2000 nodes per day but with targeted crawling enabled this speed is reduced to around 100 per day. This reduced speed is due to the greatly increased number of page loads that the crawler must perform during the filtering stage. The greedy crawler would only have to perform two page loads: the friends page and the likes page. The targeted crawler has to perform on average 240 page loads (Equation 1).

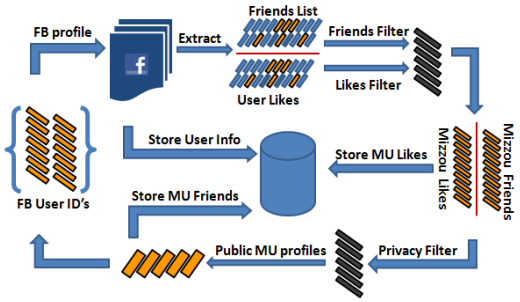


Figure 3 – Facebook Crawler

Algorithm 1 – Web Crawler

1. Login to Facebook
2. Initialize queue with seed node
3. Step 1:
4. If queue empty:
5. Go to 23
6. Pop node
7. Go to about page and grab desired information
8. Go to node like page and use modified version of algorithm in [4] to grab like information
9. Go to node friend page and grab friends list using algorithm in [4]
10. Step 2:
11. While node friend queue not empty:
12. Pop node
13. Check privacy settings of new node
14. If private:
15. Skip node; go to Step 2
16. Check if node attends the University of Missouri
17. If does not attend:
18. Add to black list; go to Step 2
19. Add node to queue if not already inspected
20. Insert node into parent node's friend list
21. Go to Step 2
22. Go to Step 1
23. End

Number of Pages = Check Friends + Check Private
= N + (Percent Mizzou Friends) * N
= 200 + .2 * 200 = 240

On the check friends stage, N friends will have to be analyzed. The check private stage must check privacy status of friends that attend Mizzou. The average user has 200 friends [] and our data shows that ~20% aren't private.

Equation 1 – Page loads

4.2 Data Analysis

As noted in 4.1, among the data collected for each crawled individual was their college major and friends that attend Mizzou. We viewed the major of the individual as declaration within that particular academic community. However, because of the number of majors available we decided to compact each major into its corresponding college (e.g. Computer Science -> College of Engineering) in order to reduce this

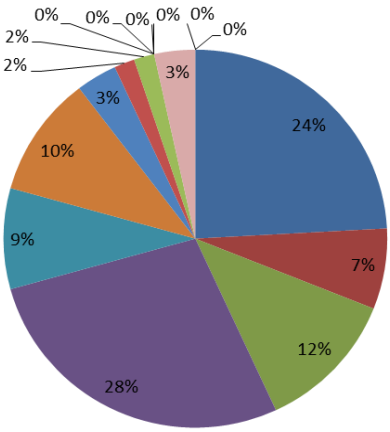


Figure 6 – EVC Top 20%

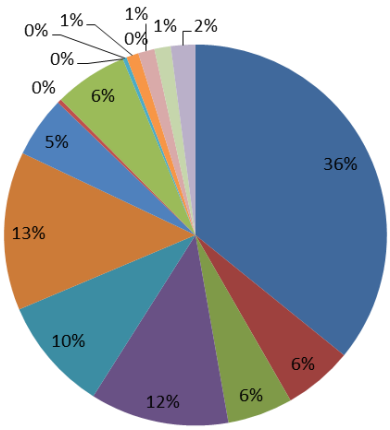


Figure 6 – Community Membership

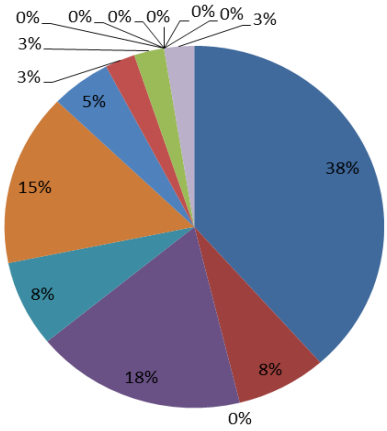
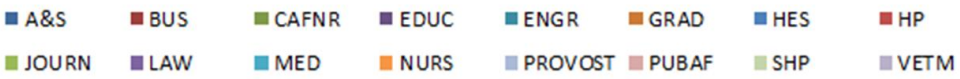


Figure 6 – PCC Top 20%



number for a clearer analysis. We first computed the percentage of membership in each college for all crawled nodes in order to have a baseline (Fig. 5). We then ran EVC on the data set and analyzed the membership for nodes within the top 20% (Fig. 4). Notably, Education majors make up a disproportionate amount of membership in the top 20%.

To understand why, we analyzed the diversity of college membership within the friendship networks of the nodes in the top 20%. Analysis showed that education majors did not have a very diverse network; on average (we need to determine this for the whole network as well), ~50% of their friends were also education majors. Other majors tended to follow the baseline (We need a graphic for this analysis). Because education majors tend to have high connectivity with other education majors, it is plausible that one highly connected education major could lead its less significant education major friends to a higher EVC score, due to the tendencies of EVC noted in 2.1. Further analysis needs to be done to ensure the correctness of this conclusion.

A similar analysis was run with PCC; we examined the top five feature vectors with the top 20% of nodes from each vector analyzed (Fig. 6). Compared to the baseline, PCC is far more comparable to it than EVC was. Because PCC analyzes the top P feature vectors of the graph, it has the ability to sample multiple areas of the network. Statistically, this should allow for a more accurate representation of the underlying network. Further analysis needs to be done to test this assumption.

EVC has been shown to give members of a somewhat small and fairly homogenous community a high score; this shows that it simply will not work as a community-centered centrality measure. PCC appears to rate nodes in manner that follows underlying network, meaning it has the same issues as EVC.

5. Framework for a New Centrality Method

Our data analysis has led us to believe that the current centrality measures are incapable of being used as community-centered measures. We propose that a new centrality measure is necessary, one that explicitly incorporates the notion of community. We propose this new measure follow the following framework:

- Resilience against isolation due to edge severance
- Number of Inter-community and Extra-community connections
- Value of community membership
- Ability to direct resources to (influence) other nodes

6. Conclusion

Our data analysis has found that EVC and PCC are inadequate for use as community-centered measures, showing the necessity of a new measure that explicitly considers community when computing a score. We have proposed a framework of principles by which this measure could be developed. Furthermore, we have implemented software that we feel will be valuable to those in the field of network analysis.

7. Future Work

We need to verify the integrity of our data analysis by utilizing a vastly larger dataset (currently the dataset is 297 nodes). If analysis on a larger data set yields similar results, we can proceed to the future work as proposed below.

The main goal of our future work is to eventually develop a new centrality measure that takes account the community membership of a node. To do this, we will use the previously mentioned principles in section 5. After developing our new centrality measure, we will need larger data sets to test it on. Thus, we would like to gather more data from Facebook and eventually other social networks such as Twitter, LinkedIn, and Google+. For our web crawler, we would like to eventually implement the support for parallel crawling and hope to eventually release it for other researchers to use.

8. Acknowledgements

This work was sponsored by the National Science Foundation of the United States.

9. References

[1] M. U. Ilyas, H. Radha, "A KLT-inspired Node Centrality for Identifying Influential Neighborhoods in Graphs", 2010 44th Annual Conference on Information Sciences and Systems, IEEE, Princeton, NJ, March 2010, pp. 1-7.

[2] M. U. Ilyas, H. Radha, "Identifying Influential Nodes in Online Social Networks Using Principal Component Centrality", The 2011 IEEE International Conference on Communications , IEEE, Kyoto, June 2011, pp. 1-5.

[3] J. Yang, J. Leskovec, "Defining and Evaluating Network Communities based on Ground-truth", The 12th IEEE International Conference on Data Mining , IEEE, Brussels, Dec. 2012, pp. 745-754.

[4] Z. Xiao, B. Liu, H. Hu, T. Zhang , "Design and Implementation of Facebook Crawler Based on Interaction Simulation", The 11th IEEE International Conference On Trust, Security And Privacy In Computing And Communications , IEEE, Liverpool, June 2012, pp. 1109-1112.

[5] N. Salamos, E. Voudigari, T. Papageorgiou, M. Vazirgianni,, "Design and Implementation of Facebook Crawler Based on Interaction Simulation", 2012 IEEE International Conference on Green Computing and Communications, IEEE, Athens, Nov. 2012, pp. 368-371.

[6] J. Xie, S. Kelley, B. Szymanski, "Overlapping Community Detection in Networks: The State-of-the-Art and Comparative Study", ACM Comput. Surv. 45, 4, Article 43 (August 2013), 35 pages.

[7] http://snap.stanford.edu/data/ca-GrQc.html

[8] www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook