Comparing Generalizations of Unweighted Network Measures

Sherief Abdallah
University of Edinburgh (Fellow), UK
British University in Dubai, UAE
sherief.abdallah@buid.ac.ae

Habab Musa
British University in Dubai, UAE
60006@buid.ac.ae

ABSTRACT

Mining and analyzing complex networks have received significant attention in recent years due to the explosive growth of social networks and the discovery of common patterns that govern wide-range of real world networks. Most of the work that analyzed social networks relied on computing measures that capture some aspect of the network structure, such as the node degree. Unweighted network measures (the class of network measures that ignore edge weights) received the bulk of researchers' attention, due to their simplicity, intuitiveness, and the relative ease of computation. This situation motivated recent generalizations of unweighted network measures that take edge weights into account. With several possible generalizations for different unweighted measures, the issue of comparing these generalizations and quantifying their effectiveness becomes increasingly important. Up until now, such comparison relied primarily on visual inspection of different plots and informal articulation on how a particular generalization is more informative than the original unweighted measure. We investigate here this issue and provide an objective automated methodology for comparing different generalizations against each other and against the original unweighted measure. As a proof of concept, we present the first comparative study of two different generalizions of the degree measure over 7 datasets.

1. INTRODUCTION

Mining and analyzing complex networks have received significant attention in recent years due to the explosive growth of social networks and the discovery of common patterns that govern wide-range of real world networks [14, 3, 7, 5, 10, 6, 13]. Significant part of the research on mining complex networks uses *network measures*, which are the functions that summarize the network structure to simpler numeric values. These measures are generally classified into two main classes: measures that ignore edge weights and focus primarily on the structure of the graph, which we call unweighted measures, and measures that take edge weights into account (in addition to the structure), which we call weighted measures.

Unweighted measures received the bulk of researchers' attention, due to their simplicity, intuitiveness, and the relative ease of computation. Such an attention resulted in several influential findings such as the small world (relied on the clustering coefficient) [14] and the power-law (relied on the degree distribution)[3, 6]. Despite their popularity, unweighted measures ignore important network information: the weights. This situation motivated the search for generalizations of unweighted measures that take weights into account [4, 2, 12, 1, 11]. With different generalizations for different measures, the issue of comparing these generalizations and quantifying their effectiveness becomes increasingly important. In particular, we would like to answer the following questions:

- Given a generalization of an unweighted measure, does the generalization provide more information than the original unweighted measure?
- Does the effectiveness of a generalization depend on the dataset (the network) involved?
- Given two generalizations, does one of the generalizations dominate the other generalization in terms of the information it provides? is this consistent across different datasets?

Having answers to these questions provide important guidelines for researchers who wish to use generalizations of unweighted network measures. Up until now, answering the above questions relied primarily on visual inspection of different plots and informal articulation on how a given generalization is more informative than the original unweighted measure [4, 2, 1, 11]. Furthermore, and although several generalizations were proposed (Section 2 provides a brief review), no experimental comparative study has been conducted between any of these generalizations.

This work presents the first comparative study between two different generalization schemes of unweighted network measures.[1] We propose a methodology for automatically and objectively assessing the effectiveness of different generalizations through node classification, as we describe in detail in Section 3. We conduct a comparative study between two state-of-the-art generalizations [1, 11], over 7 different

[1]By *generalization scheme* we mean a generalization technique that can generalize more than one unweighted network measure using the same idea.

MLG '11 San Diego, CA, USA

datasets, and using two different classification algorithms. We show that one generalization dominates the other generalization. Surprisingly, we also show that one generalization performs worse than the original unweighted measure, which raises serious questions regarding the usefulness of such generalization.

The following section gives a brief overview of the previous work in generalizing unweighted network measures before presenting our methodology for comparing different generalizations.

2. BACKGROUND

The last few years have witnessed several attempts to generalize different unweighted network measures. Here, we briefly review some of these attempts, with particular focus on the two generalizations that we will compare in this study, the C-degree and the α-degree.

2.1 C-degree: a generalization that captures the focus of interaction

A recent method for generalizing unweighted network measures captured how focused are the interactions of a given node. As an example, consider the generalized degree measure using this approach, the **C-degree**, which Figure 1 illustrates. A node on the boundary has an out degree of 1, while an internal node has an out degree of 2. Intuitively, however, only one of the internal nodes is fully utilizing its degree of 2 (the one to the left), while the other node (to the right) is focusing its interaction on only one neighbor. The C-degree measure captures this and shows that the internal node to the left has a C-degree of $c(\{0.5, 0.5\}) = 2$ while the other internal node has a C-degree of $c(\{0.9, 0.1\}) = 2^{H(0.9,0.1)} = 1.38$.

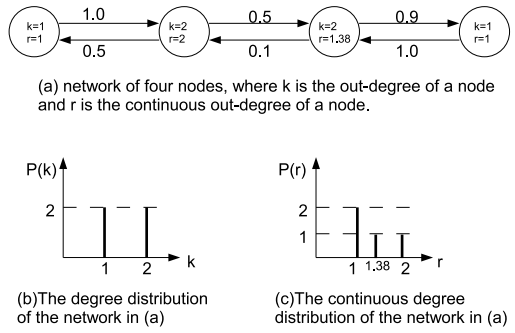

Figure 1: Example weighted network of four nodes, comparing the (discrete) degree against the C-degree. The degree distribution illustrates the benefit of taking weights into account in distinguishing nodes [1].

More formally, the *C-degree* of a node i in a network is $c(i)$ is defined as

$$c(i) = \begin{cases} 0 & \text{if } i \text{ is disconnected} \\ 2^{\left(\sum_{e \in E_i} \frac{w(e)}{s(i)} \log_2 \frac{s(i)}{w(e)}\right)} & otherwise \end{cases}$$

Where the set E_i is the set of edges incident to node i, the function $w(e)$ returns the weight on edge e, and $s(i) = \sum_{e \in E_i} w(e)$ is the strength of node i. The same idea can be applied to other unweighted network measures such as the clustering coefficient, the dyadicity and the heterophilicity [1], but here we focus primarily on the degree measure. The C-degree has three useful properties. The C-degree of a node is maximum and equals the traditional discrete degree when all the weights incident to the node are equal. The C-degree of a connected node is minimum and equals one if all edges incident to the node have zero weights except one edge that has a weight greater than zero. And finally, everything else being equal, a node with more uniform weights incident to it (less focused interaction) has higher C-degree than a node with less uniform weights incident to it.

2.2 α-degree: a generalization that mixes the strength and the degree

Another recent generalization of the unweighted degree measures mixed a node's degree with the node's strength using a tuning parameter α [11]. The generalization of the degree, the α-degree, is defined as

$$a(i) = k(i)^{1-\alpha} \times s(i)^{\alpha}$$

Where *k(i)* is the traditional degree of node i [11]. Unlike the C-degree approach, the connection between the α-degree and the original degree depends primarily on the tuning parameter α, not on the particular values of edge weights. When $\alpha = 0$, the α-degree reduces to the traditional degree regardless of the actual weights. When $\alpha = 1$, the α-degree reduces to the strength. For other values of α there is no clear guidelines for setting the α parameter. We evaluate the α-degree generalization using the two settings that were used in the original paper: $\alpha = 0.5$ and $\alpha = 1.5$ [11].

2.3 Other generalizations

There have been several attempts to generalize specific unweighted measures. The weighted clustering coefficient [4] was an attempt to generalize the clustering coefficient. The generalization relied on an alternative definition of the clustering coefficient that used triplets [14]. A triplet connected to a node is a subgraph containing the original node in addition to two other connected neighbors. The intuition behind the weighted clustering coefficient for node i is to weigh every edge between two of its neighbors, j and k, using the weights on edges (i, j) and (i, k). A recent attempt to generalize the clustering coefficient used the ratio between the total *value* of closed triplets and the total value of all triplets [12]. The authors proposed four functions to evaluate (summarize) weighted triplets: the arithmetic mean, the geometric mean, the minimum, and the maximum. The ensemble approach [2] provides a methodology for generalizing almost all unweighted network measures. The first step of the method was to normalize edge weights to ensure all weights are between 0 and 1. The next step was to randomly generate an ensemble of unweighted networks from the original weighted network, where the weight of an edge represented the probability of generating the edge. The final step was to compute the generalized unweighted measure as the average of the unweighted measure for each network in the ensemble.

The following section describes in detail the methodology we have used in our evaluation.

3. METHODOLOGY

We propose here the use of within-network classification [9] to quantify the informativeness of a generalization. The methodology is illustrated in Figure 2. We start with a real-world network of labeled nodes, where each node has a class label. We then compute for each node the following network features: the degree (D), the strength (S), the α-degree for $\alpha = 0.5$ and $\alpha = 1.5$ ($\alpha 0.5$ and $\alpha 1.5$), and the C-degree (C). We then compare the informativeness of network measures against each other in pairs. For example, we compare the accuracy of classifying nodes using only the C-degree against using only the traditional degree. To increase the reliability of the comparison, we use cross-validation with significance statistical t-test. That is, we split the dataset into 10 folds using cross validation, compute the difference in accuracy between the two paired measures for each fold, and finally compute the statistical significance using the t-test.

We have studied 7 labeled datasets that are available through the NetKit-SRL tool [9]. The datasets are from 3 different domains, 4 of which represent university websites (university of Texas, University of Washington and University of Wisconsin), 2 other datasets are extracted from news articles and represent relationships between industrial companies of varied industrial sectors, and the seventh dataset is from the Internet Movie Database (IMDb) website [9]. Before presenting our analysis, Table 1 illustrates some statistics about the 7 datasets: the number of nodes in each dataset, the number of class labels, the number of edges (links), the percentage of un-weighted edges (edges with weight equal one) and the percentage of weighted edges. From Table 1, we can see that the 7 datasets provide networks that vary in their properties. The Industry data sets have the largest numbers of nodes and the lowest numbers of links (between nodes) over all other data sets, except Washington-link1. WebKB data sets have the lowest numbers of nodes compared to Industry and IMDb datasets, in contrast WebKB have very large numbers of links compared to the number of links in industry data sets, while the IMDb dataset has the largest number of links over all data sets. Also from the table we can see that the majority of links in all data sets are un-weighted (ranging from 59.63% to 87.36%).

It is worth noting that we have extended the NetKit-SRL tool in order to compute the two generalized degree measures: the α-degree and the C-degree (the original tool could only compute traditional network measures). Two classifiers were tested using the WEKA tool [8]: the logistic regression classifier and the decision tree classifier (J48). The following section presents and discusses the results we have obtained.

4. RESULTS AND DISCUSSION

Table 2 below summarizes the difference in classification accuracies between different pairs of network measures on the 7 datasets, using the logistic regression classifier. The number in each table cell represent the average (mean) difference over 10 folds. A dash in a table cell refers to an insignificant difference in classification accuracy for a given pair of measures for a particular dataset. So for example, the first row shows that the mean difference in accuracy when classifying network nodes using only the degree minus the accuracy when classifying the same nodes using the strength is insignificant in case of the IMDB-all dataset. The difference in accuracy for Ind-pr dataset is in favor of the strength (negative difference), which is statistically significant yet quantitatively small (only 2.2% difference in accuracy).

Given a generalization of an unweighted network measure, does the generalization provide more information than the original unweighted measure? From Table 2, neither the C-degree nor the α-degree results in consistently higher accuracy against the unweighted degree. This can be seen from the pairs (rows) CD - D, $\alpha 0.5$ D - D, and $\alpha 1.5$ D - D, where the differences are mixed between negative and positive differences. The J48 classifier in table 3 shows similar results, but with better performance (accuracy) of the *C-degree* generalization measure. The *C-degree* clearly outperforms the traditional strength measure. Surprisingly, the *α-degree* generalization (for both $\alpha = 0.5$ and $\alpha = 1.5$) shows worse performance than the traditional unweighted degree and strength for both logistic and J48 classifiers, except in Industry-pr dataset where the two *α-degree*s outperform the unweighted degree using logistic regression (yet with small difference in accuracy of 1.8% to 2.2%). This finding raises a question regarding the usefulness of this generalization.

Given two generalizations, does one of the generalizations dominate the other generalization in terms of the information it provide? Is this consistent across different datasets? There is no clear winner when the C-degree is compared directly to the α-degree. This can be seen from the pairs (rows) CD -$\alpha 0.5$ D, and CD - $\alpha 1.5$ D in both Table 2 and Table 3 (again the differences are sometime positive and sometime negative). However, if we compare the best accuracy for each approach (across classifiers) we notice that the *C-degree* outperforms the *α-degree* consistently (for all datasets, C-degree either outperforms the α-degree or no significant difference. Furthermore, when the C-degree is complemented with the strength (i.e. the classification is done using both the C-degree and the strength as the node features), the combination consistently outperforms the α-degree and the traditional discrete degree (even when combined with the strength). The last three rows of both Table 2 and Table 3 illustrate this point. On the other hand, the comparison between α 0.5 and α 1.5 degrees shows no significant differences using both logistic regression and J48 classifiers, except for Texas where the α 0.5 degree outperforms the α 1.5 degree using the two classifiers.

Does the effectiveness of a generalization depends on the type of the dataset (the network) involved?

Results in tables 2 and 3 emphasize that the type of the dataset involved affects the effectiveness of generalizations. For example, the *C-degree* outperforms the α 0.5 degree in the two Industry datasets and Washington-link dataset using logistic and J48, unlike other WebKB datasets. Moreover, the *C-degree* outperforms the *α-degree* in both Industry-yh and Washington-link using logistic and J48 classifiers. The α 0.5 degree outperforms the α 1.5 degree in Texas dataset using logistic and J48 classifier, while the situation does not hold for and other dataset. Moreover, the α 1.5 degree maintains lower performance than the traditional unweighted degree inTexas and Wisconsin only, using logistic and J48. In general, the Industry datasets exhibits more consistency in results, using the two Weka classifiers, than the WebKB except the Washington-link (WebKB results are mixed between negative and positive). The IMDb dataset maintains statistically insignificant results for both logistic regression and J48 classifier.

We believe the inconsistency in WebKB datasets behav-

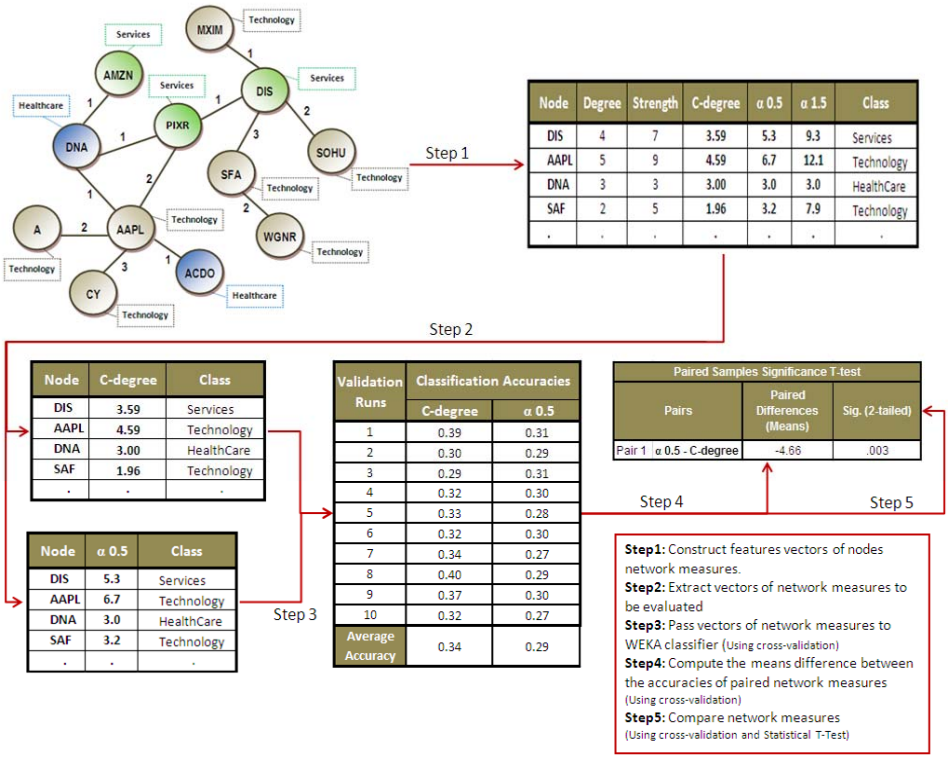

Figure 2: An illustration of the evaluation methodology.

Data Key	IMDB	Industry		WebKB			
	Imdb-all	Ind.-pr	Ind.-yh	Texas-cocite	Washing.-cocite	Washing.-link	Wiscon.-cocite
Number of Nodes	1441	2189	1798	338	434	434	354
Number of Class Labels	2	12	12	6	6	6	6
Number of Edges	51481	13062	14165	32988	30462	1941	33250
% of Un-weighted Edges	87.36%	74.74%	64.48%	74.29%	59.63%	87.27%	78.95%
% of Weighted Edges	12.64%	25.26%	35.52%	25.71%	40.37%	12.73%	21.05%

Table 1: Statistics of the 7 data sets

ior reflects the insignificant effect of edge weights. This may be due to the method of assigning weights to co-citation between web pages. The weight of link between two web-pages P_x and P_y equal to the multiplication of the total number of hyperlinks from P_x to P_z and the total number of links from P_y to P_z [9]. We suspect this multiplication of weights contribute to the less meaningful weights (not clear what is the intuition). In contrast, the industrial domain shows more consistency in results compared to WebKB. In the industrial datasets, the method of establishing links between two companies and assigning weights to links shows clearer representation of relationship between nodes than what found in other domains. A link between two industrial companies is placed if they appeared together in a news story or a press release and the weight of a link represents the number of times such co-occurrences found in the complete corpus [9]. The low base accuracy for Industry datasets (around 28%) is also worth noting and may contribute to the improvement in performance (bigger space for improvement).

A natural question then is how sensitive the results are with respect to the classifier used. Table 4 illustrates the differences in the average accuracy between the J48 classifier and logistic classifier for each individual network measure on each involved dataset. Again only statistically significant differences are shown. We can see that except for the ind-yh and the Washing-link dataset, the decision tree classifier (J48) outperformed the logistic classifier. In other words, it is the datasets that determines which classifier is more suitable than another classifier, rather than the specific network measures. This observation can be helpful when benchmarking future generalizations.

5. CONCLUSION

We present in this paper the first comparative study between two generalization methodologies of the node degree. We use the accuracy of classifying network nodes as the main evaluation metric. Our results show that the decision tree classifier generally outperforms the logistic regression classifier. More importantly, when the degree generalization are used individually as the classification feature, none of the degree generalizations we studied outperforms the original unweighted degree consistently over all datasets. However, one generalization (the C-degree) when combined with the strength, consistently outperforms both the other generalization as well as the original degree.

Our proposed methodology for comparing generalizations,

Measures Pairs	IMDb	Industry		WebKB			
	IMDb-all	Ind.-pr	Ind.-yh	Texas-cocite	Washing.-cocite	Washing.-link	Wiscon.-cocite
D - S	-	-2.2%	-	-	-	-	-
CD - S	-	4.2%	7.40%	-	-9.5%	7.90%	-
CD - D	-	6.4%	7.70%	-4.8%	-9.0%	-	-6.9%
α0.5 D - D	-	1.8%	-	-	-	-	-
α1.5 D - D	-	2.3%	-	-6.2%	-	-	-9.4%
CD - α0.5 D	-	4.7%	7.4%	-	-8.5%	8.1%	-5.5%
CD - α1.5 D	-	-	7.4%	-	-5.1%	9.2%	-
α0.5 D - α1.5 D	-	-	-	5.6%	-	-	-
CD&S -D&S	-	6.6%	7.60%	-	-	-	-
CD&S - α0.5 D	-	6.9%	7.6%	-	-	8.8%	-
CD&S - α1.5 D	-	6.4%	7.6%	7.7%	-	9.9%	7.4%

Table 2: Logistic Regression: Significance Test Results

Measures Pairs	IMDb	Industry		WebKB			
	IMDb-all	Ind.-pr	Ind.-yh	Texas-cocite	Washing.-cocite	Washing.-link	Wiscon.-cocite
D - S	-	-	2.60%	10.4%	-	-	6.0%
CD - S	-	10.4%	5.10%	10.6%	-	11.50%	-
CD - D	-	10.2%	-	-	-	12.40%	-4.8%
α0.5 D - D	-	-	-2.60%	-	-	-	-
α1.5 D - D	-	-	-3.20%	-13.4%	-	-	-5.4%
CD - α0.5 D	-	5.1%	11.9%	-	-	12.4%	-
CD - α1.5 D	-	5.7%	10.8%	13.6%	-	12.4%	-
α0.5 D - α1.5 D	-	-	-	9.8%	-	-	-
CD&S -D&S	-	12.4%	6.10%	-	-	8.40%	-
CD&S - α0.5 D	-	7.0%	13.4%	-	-	9.5%	-
CD&S - α1.5 D	-	7.6%	12.3%	14.2%	-	9.4%	-

Table 3: J48: Significance Test Results

although not optimal or perfect, can provide an initial intuition as to which generalization is potentially useful when analyzing a given network. A natural future direction of this work is to extend our study to evaluate other generalization methodologies (such as the ensemble method) as well as other unweighted network measures (such as the clustering coefficient). We also would like to include more datasets in our evaluation and investigate what dataset features determine the suitability of a given generalization.

6. REFERENCES

[1] S. Abdallah. Generalizing unweighted network measures using the effective cardinality. In *Proceedings of the 3rd international workshop on Social Network Analysis (SNA), International Conference of Knowledge and Data Discovery*, pages 52–57, New York, NY, USA, 2009. ACM.

[2] S. E. Ahnert, D. Garlaschelli, T. M. A. Fink, and G. Caldarelli. Ensemble approach to the analysis of weighted networks. *Phys Rev E*, 76(1), 2007.

[3] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[4] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Science*, 101:3747–3752, Mar. 2004.

[5] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1):2, 2006.

[6] A. Clauset, C. Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *ArXiv e-prints*, June 2007.

[7] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *Comput. Commun. Rev.*, 25:251–262, 1999.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[9] S. A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. *J. Mach. Learn. Res.*, 8:935–983, 2007.

[10] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, 2006.

[11] T. Opsahl, F. Agneessens, and J. Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, July 2010.

[12] T. Opsahl and P. Panzarasa. Clustering in weighted networks. *Social Networks*, 31(2):155–163, May 2009.

[13] J. Park and A.-L. Barabasi. Distribution of node characteristics in complex networks. *Proceedings of the National Academy of Science*, 104:17916–17920, Nov. 2007.

[14] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, June 1998.

Measures Pairs	**IMDb**	**Industry**		**WebKB**			
	IMDb-all	Ind.-pr	Ind.-yh	Texas-cocite	Washing.-cocite	Washing.-link	Wiscon.-cocite
D.J48 - D.LR	-	1.8%	-	10.7%	-	-	3.4%
CD.J48 - CD.LR	-	5.6%	-5.7%	15.7%	8.7%	-	5.4%
S.J48 - S.LR	-	-	-3.4%	-	-	-	-
$CD\&S$.J48 - $CD\&S$.LR	-	4.8%	-4.1%	10.0%	-	-	-
$D\&S$.J48 - $D\&S$.LR	-	-	-2.6%	16.9%	-	-12.8%	-
α 0.5 D .J48 - α 0.5 D .LR	-	-	-3.4%	7.7%	-	-	-
α 1.5 D .J48 - α 1.5 D .LR	-	-	-4.1%	-	-	-	7.4%

Table 4: J48-logistic regression: Significance Test Results