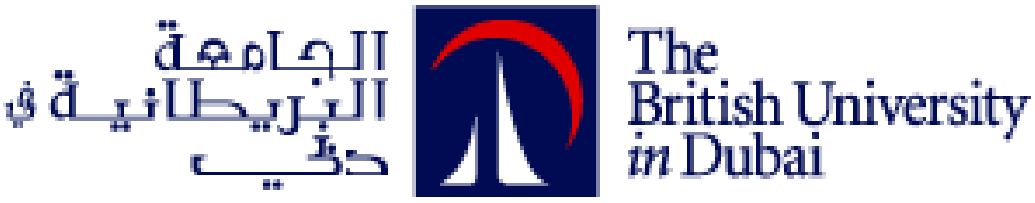


Quantitative Evaluation of Generalizations

التَـــقييم الكَـمّي للـتَعميمَـــات

By
Habab Musa Mohamed Ahmed Mohamed

Dissertation submitted in partial fulfillment of
MSc in Information Technology (Knowledge and Data Management)

Faculty of Informatics

Dissertation Supervisor

Dr Sherief Abdallah

May 2011

DISSERTATION RELEASE FORM

Student Name	Student ID	Programme	Date
Habab Musa Mohamed Ahmed Mohamed	60006	MSc in Information Technology (Knowledge and Data Management)	2 June 2011

Title

Quantitative Evaluation of Generalizations

I warrant that the content of this dissertation is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that one copy of my dissertation will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make that copy available in digital format if appropriate.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my dissertation for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

Signature

Habab Musa Mohamed Ahmed

To my great mother, to my father's soul, to my sister and brothers for their unconditional love, support and prayers throughout the stages of this thesis. To all challenges that shaped my personality.

Abstract

Inspired by the explosive growth of complex networks and the extraction of common patterns from varied complex networks' features, mining and analyzing networks have become a recent field of significant interest for many researchers with the primary focus on *network measures*. The relative ease of computation of unweighted measures leads them to be widely used in analyzing real world networks, although they ignore important network information: the weights. Despite many real world networks arise in the form of weighted networks, a few number of network measures take the weights into account. From this prospective, the last few years have witnessed the attempts of some researchers to generalize different unweighted network measures. With several possible generalizations for different measures, the issue of evaluating these generalizations and quantifying their effectiveness becomes increasingly important. Up until now, such generalizations 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.

In this thesis, we provide a comparative automated methodology for quantitative evaluation of different generalizations of unweighted degree measure. We conduct a comparative study between two state-of-art generalizations, the unweighted degree generalization based on *effective cardinality* [1] and the α-*degree* generalization [23], based on the quantitative evaluation of their productive power of classifying networked nodes. We show that some generalizations of unweighted degree measure outperform other generalizations and even the original degree measure. We study the effect of the type of the network involved and classifier used on the effectiveness of generalizations.

ملخص الرســـالـــة

أدى النمو الهائل للشبكات المعقدة (Complex Networks) وأساليب استخراج الأنماط المشتركة من الميزات المختلفة لهذه الشبكات إلى استحوذ مجال تنقيب وتحليل بيانات الشبكات المعقدة (Mining and Analyzing Networks) في العالم الحقيقي على اهتماماً كبيراً لدى الكثير من الباحثين مع التركيز بشكل أساسي على قياسات الشبكة (Network Measures). هذا وقد أسهمت السهولة النسبية لحساب القياسات غير الموزونة للشبكات (Unweighted Network Measures) في انتشار استخدامها على نطاق واسع في تحليل الشبكات المعقدة في العالم الحقيقي، على الرغم من إهمالها لأوزان الصلات بين العناصر في الشبكات (Weights) و التي تعتبر معلومة مهمة عن الشبكات.و على الرغم من أن العديد من الشبكات في العالم الحقيقي تنشأ في شكل شبكات موزونة (Networks Weighted) ،فإن عدد قليل من قياسات الشبكة تأخذ الأوزان الخاصة بالصلات بين عناصر الشبكة بعين الإعتبار. ومن هذا المنطلق، فقد شهدت السنوات القليلة الماضية محاولات بعض الباحثين لتعميم القياسات غير الموزونة للشبكات (Generalization of Unweighted Network Measures) بحيث يتم أخذ أوزان الصلات في الحسبان. ونظراً لوجود العديد من التعميمات المتنوعة للقياسات المختلفة للشبكات، يعتبر تقييم هذه التعميمات وقياس مدى فعاليتها ذو أهمية بالغة على نحو متزايد. هذا وتعتمد مقارنة التعميمات إلى الآن بشكل أساسي على الفحص و التقييم البصري للرسومات و صيغ informal articulation حول كيفية صياغة تعميمات أكثر فعالية من القياس الأصلي غير الموزون.

تقدم هذه الرسالة منهجية للمقارنة الآلية (Comparative Automated Methodology) للتقييم الكمي للتعميمات المختلفة للقياس غير الموزون لدرجة الشبكة (Unweighted Degree Measure). هذا وقد قمنا بإجراء دراسة مقارنة بين تعميمين لقياس درجة الشبكة، استناداً إلى التقييم الكمي لقوتهما في تصنيف العناصر ضمن الشبكة الواحدة، وهما تعميم درجة الشبكة باستخدام الأصول الفعالة (effective cardinality) [1] والتعميم باستخدام الدرجة ألفا (α-Degree) [23]. و قد أثبتنا من خلال الدراسة و البحث أن بعض التعميمات غير الموزونة للقياس درجة الشبكة تفوق غيرها من التعميمات، كما تفوق قياس الدرجة الأصلي. كما قد قمنا بدراسة تأثير نوع الشبكة (Network Type) والمصنف (Classifier) المستخدم في فعالية التعميمات.

Acknowledgement

This thesis is of a great value for me. In fact, this thesis is a product of myself and my interdependence with others. Without the support of all those who help me, it would not have been possible to complete this thesis. It is a pleasure to have the opportunity to express my sincerest gratitude and respect to them here.

For my academic achievements, I would like to express my sincerest gratitude to my supervisor, Dr.Sherief Abdallah, for his guidance and inspiration during my work on this thesis. His very rich knowledge, understanding and encouragement have been a great support to pass the obstacles during the thesis research and writing. I would like to express my gratitude to Dr.Sofus Attila Macskassy, the author of NetKit-SRL network learning toolkit, for this kind response to my questions about the toolkit java components in the initial stage of this work.

On a social note, I would like to express my unlimited thanks to my great mother, my sister Rihab, my brothers Yousif and Mohamed, and my aunt Alawya Alhala for their love and unlimited support. Without them I would have lost motivation and focus.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Habab Musa Mohamed Ahmed Mohamed*)

Contents

List of Figures

List of Tables

Chapter 1

Overview

1.1 Introduction

Inspired by the explosive growth of complex networks and the discovery of the heterogeneous complex networks' features, mining and analyzing networks have become a recent field of significant interest for many researchers with the primary focus on *network measures*. Network measures are the core of mining complex networks that are capable of expressing important complex network features. In past studies, unweighted network measures are of wide popularity in analyzing complex networks due to their relative ease of computation and intuitiveness; although they focus on network structure and ignore the weight assigned to each link (links between elements are usually considered as binary states, either present or absent). Such studies resulted in important complex networks findings such as the power-law (introduced based on the degree distribution) [3, 7] and the small-world (introduced based on the clustering coefficient) [28]. The situation differs in weighted network measures which take weights into account along with the network structure, where the weights are of great importance and are assigned to links proportional to the intensity of the connections between network elements. This situation motivated the researchers to develop different generalizations of unweighted network measures to take weights into account to capture the richness of information in connections weights [1, 2, 4, 24]. With several possible generalizations of different network measures, the issue of evaluating these generalizations and quantifying their effectiveness becomes increasingly important. Up until now, such generalizations 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. To our knowledge, no any experimental comparative study has been conducted to evaluate quantitatively different generalizations.

In this thesis, we provide an automated comparative methodology for comparing quantitatively generalizations of unweighted degree measures. We conduct a comparative study between two state-of-art generalizations, the unweighted degree generalization based on *effective cardinality*, the *C-degree* [1], and the *α-degree* generalization [23]. We will use the improvement in the classification accuracy as an indication of the additional information provided by the generalization. The two generalizations will be evaluated using different datasets from different domians and using different classifiers. We show that some generalizations of unweighted degree measure outperform other generalizations and even the original degree measure, while other generalizations show lower performance. We present interesting findings related to the effect of combining some generalizations with original measures. We study the effect of the type of network involved on the effectiveness of generalizations as well as the sensitivity of the results with respect to the type of classifier used.

1.2 Problem Statment

A serious need of guidelines for the effectiveness of different methodologies of generalizations of unweighted network measures arises due to the explosive use of network measures in analyzing complex networks. The issue of comparing these generalizations and quantifying their effectiveness becomes increasingly important. Up until now, such generalizations 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. The key objective of this thesis is to provide an automated comparative methodology for comparing quantitatively two state-of-the-art generalizations of unweighted degree measures.

To accomplish our objectives, we developed an automated comparative methodology for comparing quantitatively two state-of-the-art generalizations of unweighted degree measures and assessing how useful and informative these generalizations are, when compared with the original unweighted network measure and against each other. We conduct an extensive experimental comparative study between two generalizations of unweighted node degree measure, the *C-degree* and the *α-degree*, and against their original unweighted measure using different datasets from different domains and using different classifiers. We use the accuracy of classifying networked nodes as the main evaluation metric.

1.3 Research Questions

In this thesis, we address the following research questions:

- Given a generalization of an unweighted network measure, does the generalization provide more information than the original unweighted measure?
- 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?
- Dose the effectiveness of a generalization depends on the type of the dataset (the network) involved?

1.4 Contribution

The contributions of this thesis are:

- Providing an automated comparative methodology for comparing quantitatively two state-of-the-art generalizations of unweighted degree measures.
- Conducting an extensive experimental comparative study between two different generalizations of unweighted degree measures and against the original unweighted degree measure.
- An extensive study of the effect of involving different types of networks and using different classifiers on the effectiveness of generalizations.

1.5 Scope

In this thesis, we focus on experimental comparative analysis of two state-of-the-art generalizations of unweighted degree measures. We restrict our analysis to weighted and undirected real world complex networks. It is our aim that the contributions made in this thesis to provide important guidelines for researchers who wish to use different generalizations in mining and analyzing real world complex networks.

1.6 Organization of Thesis

The remaining chapters are arranged as follows: Chapter 2 presents a literature review about networks, different classes of network measures and a review of recent generalization methodologies. Chapter 3 provides basic background of within-network classification, node-centric network learning framework and the modular network learning NetKit-SRL Toolkit. At the end of Chapter 3, we present brief details about the datasets that will be involved in our experimental study. Chapter 4 demonstrates our proposed automated comparative methodology for comparing quantitatively generalizations of unweighted degree measures. Chapter 5 is about results analysis and discussion, it presents the datasets statistics, the experimental setup and the analysis of involved datasets. At the end of Chapter 5, we present discussion illustrates how our analysis of the results collected from our comparative methodology answers the research questions we raised in section 1.3. Finally, in Chapter 6, conclusion, ideas for future researches and enhancements are highlighted.

Chapter 2

Literature Review

This chapter first presents a brief background about networks in general, different types of networks and mining complex networks. Then it provides a brief background about different classes of network measures that will be used in this study. Finally, we survey the two recent attempts to generalize unweighted degree measure that will be evaluated in this comparative study along with other researcher's attempts of generalizations.

2.1 Networks

A network (graph in discrete mathematics) is a set of elements (nodes or vertices) connected by a set of links (edges or ties) in a system. Complex networks are networks with special characteristics such as scale-free [3], small-world effect [28], community structure and mixed patterns.

The applications of complex networks in real world arise everywhere around us in systems that are of fundamental importance to modern societies and researches. Examples of the former are social networks (personal relations, collaboration, email exchange and organizational management), economy networks (trade networks, currency and tourism), technological networks (Internet, airports, railways and electric power grid), information networks (WWW and citation networks), and biological networks (protein-protein interaction networks, metabolic networks and genetic networks) [8, 9].

Table 2.1 shows examples of real world application from four categories of networks with the corresponding nodes and edges for each [6, 20].

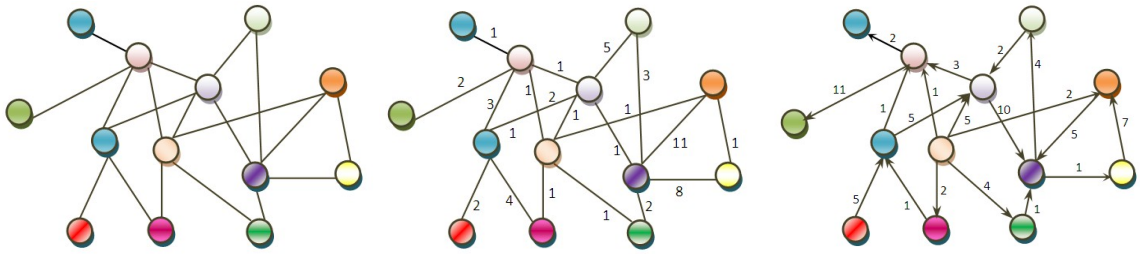
Network Category	**Network**	**Nodes**	**Edges**
Social Networks	Social groups	Individuals	Political
	E-mail exchanges network	E-mail addresses	Massages
	Industrial networks	Companies	Business relationships
	Human language networks	Words	Synonymous or syntaxes
Information Networks	Citation networks	Papers	Citations
	World Wide Web	Web pages	Hyperlinks
Technological Networks	Internet,	Hosts	Physical connections
	Airplane networks	Airports	Airline connections
Biological Networks	Metabolic pathways network	Metabolic substrates	Metabolic reactions
	Food Web	Species in an ecosystem	Predator-prey relationships

Table 2.1: Complex Networks Applications to Real Networks

The basic assumption that all edges that connecting nodes in any network are equal in terms of their capacity and contribution in the overall performance of the network may not be a valid assumption. To clarify this, here is an example, in industrial networks the nodes represent the companies and the edges are the business relationship between them. The strength of business

relationship between two companies can be measured from the weight of the edge connecting them [3, 6, 7, 9, 15, 25]. It is thus important to differentiate between two types of networks, unweighted and weighted networks:

- An unweighted network is a network with all the edges connecting nodes are considered equivalent and treated on equal footing (uniform weight of 1), as seen in figure 2.1a.
- A weighted network is a network with the edges connecting nodes have weights associated with each of them and represent the strength of connection between any two nodes in various contexts, as seen in figure 2.1b.



(a) Unweighted Network (b) Weighted Network (c) Directed and weighted network

Figure 2.1: Different Types of Networks

Networks in real world can be directed or undirected depending on the ways of interaction between nodes. In directed networks, an edge, sometimes called *arc*, has only one direction which indicates the orientation of the connection between two nodes. An edge in directed network has two components: the edge weight and the direction. As figure 2.1c shows, a node in a directed network can have a number of in-coming edges and out-going edges. For example, in e-mail exchanges network, nodes are the email addresses and the directed edges represent messages passing from one e-mail address (sender A) to another (receiver B), where passing a message from A to B, doesn't necessarily mean that a message should be passing from B to A.

In undirected networks, an edge is created when two nodes have a connection in between regardless of the direction of the connection as seen in figures 2.1b and 2.1a. A node in an undirected network has only a total number of links to other nodes, which represent the node degree. An internet movie database is an example where edge between two movies is created if they share a production company [6, 17, 22].

In this thesis we restrict our attention to undirected networks, so that we focus only on edge's weight component rather than edge's direction.

Mining networks refers to the process of extracting hidden patterns, predictive information, future trends and behaviors from large scale databases. Recently, mining and analyzing complex networks can help businesses in making knowledge-driven decisions and answering business questions through predicting the clients' future behaviors [12].

2.2 Network Measures

This section reviews some network measures which have been developed to characterize and summarize network structure into simpler numeric values and considered as the core of mining and analyzing complex networks [1].

2.2.1 Unweighted Network Measures

Most of the traditional network measures, as node's degree and clustering coefficient, are developed with primary focus on network structure without taking edges weights into account (unweighted network measures). The relative ease of computation of unweighted network measures leads them to be widely used in analyzing real world networks. Although many networks are weighted networks, an inadequate number of network measures take the weights of connections between network elements into account [24]. From the list of unweighted network measure, our concern in this thesis is on the traditional unweighted node's degree measure.

Node's Degree

The *degree* measure is a basic and an important characteristic of a node that is often used in basic studying of network characteristics [23]. The degree of a node is equal to the number of other nodes connected to it. In directed networks, which are out of the scope of this work, nodes have both an out-degree and in-degree, which are the numbers of out-going and in-going edges respectively. Node's degree suffers from two main drawbacks: it ignores the weights that are assigned to edges between nodes and its discrete nature where a neighbor is either counted in the degree or not (either the edge to neighbor is present or absent) [9, 20].

2.2.2 Weighted Network Measures

Weighted network measures are extensions of unweighted network measures to summarize weighted networks and to capture the richness of the information contained in the data by taking edges' weights into account along with network structure [1, 4, 6, 8, 24]. Since complex networks arise widely in different contexts of real world systems, they are better to be described in terms of weighted networks to signify the heterogeneity in the intensity or the capacity of connections between network nodes [6]. From the list of weighted network measures, in this work, we will focus only on the node's strength and we will not survey all other weighted network measures that are not developed based on unweighted measures (detailed reviews of these weighted measures found in [5, 6, 21]).

Node's Strength

Node's unweighted degree measure has been extended to the sum of weights attached to the edges that connect a node to others, known as *node's strength*. Node's strength can be considered as the natural way of generalizing degree as it reflects the intensity of the connection between nodes by integrating the information found in the number of edges (the degree) incident to a node with the weights assigned to each of these edges [5, 6].

Nod's strength becomes identical to the node's degree in the case of unweighted networks or in the very special case of weighted networks, when all edges incident in a node are of weights equal to 1. For example node G in figure 2.2 has identical strength and degree as all of edges incident to G are of weight equal to 1. However, node's strength doesn't ensure partial ordering among sets of weighted nodes. For example in figure 2.2, nodes A, E and F have the same strength, but node A is more involved in the network as it is connected to twice as many nodes as each of nodes E and F [1, 23].

From the brief review above and as will be seen in the coming survey of researches attempts of generalization of unweighted network measures, it is very important to involve both node's unweighted degree and strength in network analysis and mining as both reflect the level of contribution of a node in the overall performance of the network [23].

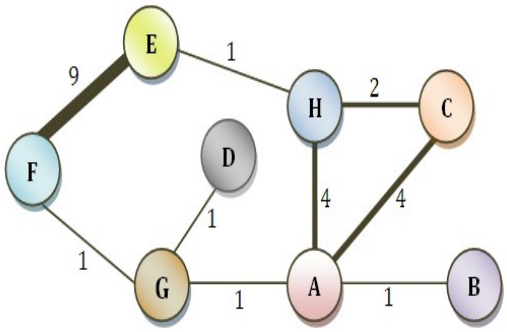


Figure 2.2: A Network With 8 Nodes and 9 Weighted Edges

2.2.3 Generalized Unweighted Network Measures

The last few years have witnessed the attempts of some researchers to generalize different unweighted network measures. Here, we briefly review some of these attempts, although we are not going to involve all of them in our experimental study.

The Newly introduced methodology for generalizing all unweighted network measures is developed through generalization of the cardinality of some subsets of edges to take weights into account. This methodology defines "*The effective cardinality, a new metric that quantifies how many edges are effectively being used, assuming that an edge's weight reflects the amount of interaction across that edge*" [1]. As the cardinality is the heart of many unweighted network measures, the generalization of the cardinality is applicable for many unweighted network measure such as: the node's degree, the clustering coefficient, the dyadicity and the heterophilicity [1].

Unlike other generalizations, this methodology ensures important properties if the weights are uniform (edges of equal weights). Effective cardinality guarantees the reduction of the generalized measures to the unweighted measures if every node interacts equally and uniformly with all its neighbors. Moreover, the effective cardinality guarantees a partial ordering among sets of weighted edges, where the generalized measure capture the strength of connections in a network in accordance with the disparity between weights [1]. One of the unweighted degree generalizations that will be used in this thesis is the continuous degree (*C-degree*) of a node which is developed based on the effective cardinality for analyzing weighted networks, the C stands for continuous. As the strength of a node becomes identical to node's degree if all edges incident to that node are of weight equal to 1, the *C-degree* also becomes identical to node's degree in the case of unweighted network or in the very special case of weighted network if all edges incident to a that node are of weight equal to 1.

The effective cardinality $c(E')$ is

$$c(E') = \begin{cases} 0 & \text{if } c(E') \text{ is empty} \\ 2^{(\sum_{e \in E} \frac{w(e)}{\sum_{o \in E'} w(o)} \log \frac{\sum_{o \in E'} w(o)}{w(e)})} & \text{otherwise} \end{cases}$$

Where E' is a subset of edges used by any network measure and $\frac{w(e)}{\sum_{o \in E'} w(o)}$ is the probability of an interaction between the node i and one of its neighbors over an edge e.

The *C-degree* of a node i is

$$c(E_i) = \begin{cases} 0 & \text{if } c(E_i) \text{ is empty} \\ 2^{(\sum_{e \in E} \frac{w(e)}{s(i)} \log \frac{s(i)}{w(e)})} & \text{otherwise} \end{cases}$$

Where E_i is the set edges incident to i , $w(e)$ is the weight of each edge incidents to i and $s(i)$ is the strength of node i.

The *C-clustering coefficient* $o(i)$ of a node i is

$$o(i) = \frac{c(E_i^N)}{MAX_i^N}$$

Where E_i^N is the set of edges between node i's neighbors and MAX_i^N is the maximum expected number of edges the neighbors of the node i.

Moreover, what distinguishes the generalization of the clustering coefficient measure using this methodology form others is that it takes the weight of edges between the neighbors of the focal node i into consideration while generalizing.

Another two recent measures that can be generalized using effective cardinality are the dyadicity and heterophilicity of the graph. These two measures summarize the network structure by representing the correlation between classes in a network [25]. The dyadicity of a graph reflects the strength of connection between nodes of the same class compared to what is expected for a random configuration and can be generalized using the effective cardinality as $\frac{c(E_{within})}{n_{within}}$, where E_{within} is the set of edges between nodes of the same class and the n_{within} is the maximum expected number of edges within nodes of the same class. For example, for a class of type (x) the $n_{within} = \frac{n_x(n_x-1)}{2}p$, where $p \equiv 2EN(N-1)$ is the connectance, representing the probability that two nodes being connected in the graph [25], and N is total number of nodes. The heterophilicity of a graph reflects the strength of connection between nodes of the different classes compared to what is expected for a random configuration and similarly can be generalized as $\frac{c(E_{across})}{n_{across}}$, where E_{across} is the set of edges between nodes of different classes and the n_{across} is the maximum expected number of edges between nodes in different classes $n_{within} = n_x(N - n_x)p$.

The *α-degree* is another recent but simple generalization methodology of the unweighted degree measure (with closeness and betweenness measures) [23]. The *α-degree* was developed based on three elements: the node's degree, the strength and a tuning parameter α.

$$\alpha\text{-}degree = k^{1-\alpha} \times s^{\alpha}$$

Where the k is node's degree and the s is node's strength [23]. Unlike the effective cardinality approach, the *α-degree* generalization depends mainly on the tuning parameter α and doesn't reflect the effect of disparity between weights, as it has no clear dependence on edges weights. Moreover, there is no decided rule for the tuning parameter setup. From *α-degree* definition, this generalization capitalize on the effect of the total sum of weights (strength) if the tuning parameter $\alpha \geq 1$, where it capitalize on the number of edges incident to the focal node if α is between 0 and 1.

In this thesis, we focus on quantifying the amount of information exposed by the continuous degree *C-degree* and comparing quantitatively its effectiveness to the *α-degree* (using the two setting that were used in the original paper: $\alpha = 0.5$ and $\alpha = 1.5$) and against the traditional unweighted degree measure.

The ensemble generalization approach is another recent attempts to develop a generalization methodology that can be applied for all unweighted network measures [2], without focusing on certain measures [4, 24]. The ensemble approach consists of two main steps. The first step in constructing generalized measures for weighted networks is to transform the weighted network to set of unweighted ensembles of edges, where the probability that the weighted edge w_{ij} between nodes i and j exists in the ensemble depends on the normalization of the edge to a quantity $p_{ij} \in [0, 1]$.

$$p_{ij} = \frac{w_{ij} - min(w_{ij})}{max(w_{ij}) - min(w_{ij})}$$

The second step after normalization is to calculate the generalization of the targeted unweighted measure, for the entire network, as the average of the unweighted measure for each network in the ensemble.

Following this ensembling procedure, any unweighted measure can be generalized to fit weighted networks. The ensemble approach of generalization is applicable for almost all unweighted measures. The evaluation of this approach on real world networks is reported using the clustering coefficient on two networks: the EU aviation passengers network and the English Language letters network. Unlike the generalized clustering coefficient in [24], the ensemble approach can be used to analyze fully connected networks and captures the topological information given by edges weights. Using probabilities in the ensemble approach leads to difficulty in distinguishing between nodes of the same degree (no partial ordering among nodes). Another drawback of probabilities is that the generalization will not be reduced to the original unweighted measure unless all weights are exactly normalized to 1, unlike the generalization approach based on effective cardinality which offer both partial ordering and reduction to unweighted measures in case of uniform weights.

One more developed method to generalize one of the unweighted network measures is the weighted clustering coefficient [4]. This generalization was developed in a study of correlation among weighted connections in complex networks architecture such as the analyzed social and large infrastructure systems. For each triplet of nodes (i, h and j), the weighted clustering coefficient C_i^w uses the weight of edges between the focal node i and each of its two neighbors, w_{ij} and w_{ih}, but unlike the *C-clustering coefficient* it ignores the weight of the edge between the pair of node's i neighbors w_{jh}.

$$C_i^w = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{(w_{ij} + w_{ih})}{2}$$

Another recent methodology of generalizing the clustering coefficient relied on the total weight of edges in the closed triplets (triangles) and the total weight of edges in all triplets [24]. The method of defining (summarizing) the weights is of primary impact on the final clustering coefficient and should be chosen, based on research nature, from the list of the proposed mathematical functions.

$$C_w = \frac{Total\ value\ of\ closed\ triplets}{Total\ value\ of\ all\ triplets} = \frac{\sum_{\tau\Delta} w}{\sum_{\tau} w}$$

Where $\tau\Delta$ represents the closed triplets (triangles) and τ represents all triplets.

The generalized clustering coefficient C_w becomes identical to the unweighted coefficient in case of unweighted networks (all weights equal to 1). A reported limitation for this method is that it uses the absolute values of weights; therefore, it is applicable only for weights on the ratio scale not the ordinal scale. To overcome this limitation, networks of ordinal weights scale should be transform to ratio scale to not affect the final value of C_w, which is not normalized unlike both the traditional and the *C- clustering coefficient*. Moreover, C_w suffers from a problem in the case of fully connected networks as it gives a C_w value of 1, unlike the traditional clustering coefficient. Another shortcoming in C_w measures and the traditional one is that they don't capture the differences between edges weights and thus don't successfully distinguish between different nodes of a same clustering coefficient equal to 1 , unlike the *C- clustering coefficient* that take weights into account to distinguish nodes that deemed indistinguishable.

In this thesis, we provide a comparative study between four network measures shown in table 2.2: two state-of-art generalizations of unweighted degree measure (the *C-degree* [1] and the *α-degree* [23]) and two traditional measures (the unweighted degree and the strength). We

Network Measure	**Formula**
Traditional unweighted Degree D	$\lvert E_i \rvert$
Traditional Strength S	$\sum_{e \in E} w(e)$
C-degree Generalization	$c(E_i)$
α-*degree* Generalization	$k^{1-\alpha} \times s^{\alpha}$ $(\alpha = 0.5 \text{ and } \alpha = 1.5)$

Table 2.2: Involved Network Measures in This Thesis

focus on comparing quantitatively the informativeness of the *C-degree* generalization and the α-*degree* generalization (using the two setting that were used in the original paper: $\alpha = 0.5$ and $\alpha = 1.5$). We also compare the two generalizations against the traditional unweighted degree and strength.

Chapter 3

Background

This chapter provides the basic background of the within-network classification, the node-centric network learning framework and its components. Then it provides the basic background of the modular and node-centric NetKit-SRL toolkit including the core routine and its five former modules. At the end, brief details about the datasets that will be involved in the experimental comparative study.

To develop our automated comparative methodology, we extend the node-centric network learning toolkit (NetKit-SRL) in order to compute the two state-of-the-art generalized degree measures, the *C-degree* and the α*-degree* , and then to pass them as additional information for data classification using different classifiers.

3.1 Within-network Classification

Our comparative methodology of different unweighted degree generalizations relies on the evaluation of the productive power of these generalizations in *within-classification* of networked nodes of missing labels. As it often the case in real world networks, the class of some interconnected nodes may be missing (unknown) and need to be recovered (or estimated). Such cases offer opportunities for extending traditional machine learning classification, which treat nodes as being independent, to classify interconnected nodes of missing information such as node class to be estimated based on the available known information of other linked nodes, known as *within-network classification*. Most of within-network classification methods are inferring the missing classes (labels) collectively based on the homophily hypothesis (the principle that interrelated nodes have higher tendency to be in the same class), while this fails for certain complex networks (*e.g.*, molecular graphs) [11, 17].

3.2 A Node-centric Network Learning Framework

The Network learning toolkit (NetKit-SRL), that we extend in this thesis, is based on node-centric framework. The Node-centric framework focuses on a single node from the networked data at a time, with appropriate choice of three main components: the collective inference (to infer node classification based on the neighbors classification simultaneously), the relational classification (to produce class labels probability for a certain node given available node's neighbors labels and local attributes as needed), and the local classifier (to initially assign labels for classes based on prior knowledge from the available local data).

3.2.1 Non-relational Classifier

Non-relational classifier or known as local classifiers is a simple machine learning method that uses available local information (attributes) of nodes of which the class labels are unknown and need to be estimated. The non-relational classifier can be used in relational learning for collective inference for assigning initial class labels (priors) [14, 17].

3.2.2 Relational Classifier

Relational classifiers make use of the relations between entities (nodes) along with the values of attributes of the related entities, or the values of local attributes in some cases, to estimate the class value for each entity [27]. To perform our comparative methodology, we customize the relational classifier module of NetKit-SRL to classify nodes of unknown labels using vectors of nodes generalized degree measures.

3.2.3 Collective Inference

Collective inference is a basic module in NetKit-SRL. Collective inference is a methodology for simultaneously inferring (classifying) the interrelated and unlabeled nodes together. Collective inference method uses both local and relational classifiers for classification of interrelated data. As a first step, collective inference uses a local classifier to estimate initial class labels (priors) for each node using local attributes only. Then, collective inference uses the resulted initial estimates to assign class probability for each node (reclassify nodes). The process of reclassifying could be repeated based on the specified number of iterations or until the class labels converge [13, 18]. Due to this iterative process of inferring, collective inference shows significant improvement in classification accuracies for interrelated nodes over standard methods that classify nodes independently and ignore relation between nodes, as shown from recent researches [16, 18, 19, 26].

To accommodate the goals of our comparative methodology, we focus only on the use of the relational classifier module from the node-centric framework.

3.3 Network Learning Toolkit (NetKit-SRL)

In this section, we introduce brief details about the Network learning toolkit that we are going to extend in this comparative study, NetKit-SRL. To achieve the goals of this thesis and to quantitatively evaluate different generalizations of unweighted degree, we will implement the *C-degree* and the *α-degree* generalizations as java components (data aggregators) to be integrated with the Netkit-SRL that could compute the traditional network measures only [17].

NetKit-SRL is a command-line toolkit written in Java 1.5 and is available online as open source. NetKit-SRL is developed based on a node-centric framework and is designed to accept the interchange of the three components from a set of different methods that are available for each the node-centric components, where a pair of any non-relational and relational classifiers can be constructed and then be combined with any selected inference method. The common platform enables NetKit user to compare different classifiers and learning methods on equal footing. Moreover, NetKit-SRL is designed to accommodate the introduction of new components; therefore, we decided to use NetKit-SRL to extend its features in order to achieve our goals.

Table 3.1 illustrates the core routine of Netkit-SRL for a given input graph $G = (V, E, X)$ where the $G^K = (V, E, x^K)$ denotes every known information about the graph G, $x^K \in X$ represents the vector of known class values for the set of vertices $v^K \in V$, and E is the set of edges between vertices(weighted and undirected) [17]. NetKit collective inference is the process

Input: G^K, V^U, RC_{type}, LC_{type}, CI_{type}
Induce a local classification model, LC, of type LC_{type}, using G^K
Induce a relational classification model, RC, of type RC_{type}, using G^K
Estimate $x_i \in V^U$ using LC. Apply collective inferencing of type CI_{type}, using RC as the model
Output: Final estimates for $x_i \in V^U$

Table 3.1: High-level pseudo-code for the core routine of the Network Learning Toolkit

of inferring values of the unknown classes $V^U = V - V^K$ for the set of remaining vertices of unknown class X^U.

The modular toolkit NetKit-SRL consists of five general modules described below:

3.3.1 Input

NetKit-SRL takes a set of three files as input data for classification:

- Schema file: A file that describes the schema of the input data and should follow the file name convention *schema.arff*.
- Nodes file: A file that describes nodes along with their corresponding attributes (in the same node type and order that is specified in schema file) and should follow the file name convention *nodetype.csv*.
- Edges file: A file that describes the edges between connected nodes in the form (*SrcNode, DestNode, Weight*) and should follow the file name convention *edgefile.rn*.

3.3.2 Local Classifier Inducer

Local classifier (non-relational), which is out of the scope of our study, is used for initializing class priors. The Netkit-SRL offers seven types of local classifiers: null (do nothing), uniform prior, class prior (the default classifier), external prior (read from external file), WEKA logistic regression, WEKA naive bayes and WEKA J48.

3.3.3 Relational Classifier Inducer

In this thesis, we use the Network-only Link-Based relational classifier (nLB), which is developed based on the link-based classifier in [16]. The basic idea behind the original nLB classifier in NetKit-SRL is that it follows two steps, the aggregation and the estimation. The original nLB generates features vector for each node by aggregating the labels of neighboring nodes using the four aggregation methods (mode, ratio, count, exist). Then the generated features vectors will be passed to logistic regression without any consideration of local attributes, unlike the link-based classifier in [16]. Finally, the unknown class labels are estimated by applying the learned model. Notably, relational classifiers in Netkit-SRL restrict the estimation process to only local neighbors of the targeted node. Moreover, Netkit-SRL offers other eleven types of relational classifiers that are described in details in the NetKit-SRL Univariate Case Study [17].

To accommodate our goal of evaluating unweighted degree generalizations against original degree measure, we extend the relational classifier in NetKit-SRL by implementing three aggregators to compute generalizations: the *C-degree*, the α 0.5 degree and the α 1.5 degree along with node's degree and node's strength aggregators (as described in section A.1). These aggregators will be used to aggregate each node with its network measures in the form of features vector. Then the generated features vectors will be passed using nLB to the specified WEKA classifier to estimate the unknown class labels.

3.3.4 Collective Inference

The Netkit-SRL uses collective inference classifier, which is out of the scope of our study and relies on relational classifier estimations in inferring the class labels for nodes of unknown class. NetKit offers four inference methods: null (no inference), gibbs sampling, iterative classification and relaxation labeling (the default method).

To accommodate the goals of our comparative methodology, which relies on the evaluation of measures productive power of classifying networked nodes, we focus only on the use of the relational classifier module from the NetKit-SRL node-centric toolkit.

3.3.5 Weka Wrapper

Two classifiers were tested in this study using a wrapper WEKA [1] module [29] introduced by the NetKit-SRL toolkit: the logistic regression and decession trees (J48). Weka classifiers in NetKit are used either as a relational classifier (to build a discriminative model based on the generated feature vector of aggregated labels of each node's neighbors) or as a non-relational classifier, which is out of the scope of our study.

The configurations of all used classifiers from WEKA are available in Weka.properties file,as shown in table 3.2, where the Toolkit enables the user to add any additional WEKA classifier that uses WEKA API to be involved in the NetKit-SRL after declaring them in properties file.

logistic.class=weka.classifiers.functions.Logistic
leastsquares.class=weka.classifiers.functions.LeastMedSq
naivebayes.class=weka.classifiers.bayes.NaiveBayesMultinomial
j48.class=weka.classifiers.trees.J48

Table 3.2: Default WEKA Configuration

In this thesis, we use both WEKA Logistic Regression and Decision Trees (J48) for classification experiments in order to study the effect of using different classifiers on generalizations effectiveness and to evaluate the sensitivity of the collected classification results in this comparative methodology to the used WEKA classifiers.

3.4 Benchmark Datasets

In this thesis we involve 7 benchmark data sets from three different domains that have been the subject of prior studies in machine learning and already used in the NetKit-SRL Univariate Case Study [17].

3.4.1 IMDb

The first dataset is from the Internet Movie Database (IMDb)[2] for movies released between 1996 and 2001 in USA with the goal of building models for movies revenue classification by estimating whether the opening weekend box-office receipts exceeded 2 million $ [19]. The movies are labeled into 2 categories: high and low revenues.

The method of establishing links between movies in this datasets is: two movies are linked if they share one or more production companies. The weight of an edge between two movies represents the number of production companies two movies share. Moreover, the sequential aspect in releasing movies is ignored while designing the dataset and may lead to movie in the training set to be released after movies in the test set.

[1] NetKit-SRL uses Weka version 3.4.2.: available at http://www.cs.waikato.ac.nz/ml/weka/.
[2] See http://www.imdb.com.

3.4.2 WEBKB

The Second domain is based on the WebKB Project[3] [10]. We use 4 datasets from computer science department in universities websites (University of Texas, University of Washington and University of Wisconsin). The 4 dataset that we use from NetKit case study is manually labeled into 6 categories: course, department, faculty, project, staff, and student. Where NetKit case study considers two different classification problems: the six-class problem and the binary-class classification for predicting students' pages from others.

In the WebKB datasets, two web pages are linked by co-citation, where the weight of a 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 by the total number of links from P_y to P_z [17].

3.4.3 Industry

The third domain is the Industry domain with 2 datasets that represent the relationship between industrial companies from varied industrial sectors and are extracted from news articles. The 2 datasets that we use from NetKit case study are labeled into 12 categories: Basic Materials, Capital Goods, Conglomerates, Consumer Cyclical, Consumer NonCyclical, Energy, Financial, Healthcare, Services, Technology, Transportation and Utilities.

A link between two industrial companies is placed if they appeared together in a news story or a press release, where the weight of a link represents the number of times such co-occurrences found in the complete corpus [17].

[3]The used data is from WebKB-ILP-98 data.

Chapter 4

Comparative Methododology

In this Chapter, we provide a brief description of the extension we did to Network Learning ToolKit (NetKit-SRL) along with the details of our comparative methodology for comparing quantitatively different generalizations of unweighted degree measure through the evaluation of their productive power for within-network classification.

We have extended the NetKit-SRL to compute the two generalized degree measures: the *C-degree* and the α-*degree* (with two parameter settings $\alpha = 0.5$ and $\alpha = 1.5$) using our newly developed java aggregators (for more details see aggregators implementation in Chapter A). Moreover, we developed another two aggregators to compute nodes degree and strength in the form of features vectors, where the original tool could only compute the traditional degree and strength measures for statistical purposes only [17]. We customized the Network-Only Link-Based classifier to use two WEKA classifiers, decision trees J48 and logistic regression, for classifying network nodes based on the constructed nodes' features vectors of different network measures (the details of the constructed network measures aggregators, classifiers and components configurations are in Chapter A).

Our comparative methodology to quantify the informativeness of a network measure follows the steps below:

1. We start with weighted and undirected real world network of labeled nodes as input to the NetKit-SRL toolkit, where each node has a corresponding class label (as seen in figure 4.1)

2. For each node we construct features vectors by aggregating the node and its network measures (Node's degree D, Node's strength S, Node's *C-degree*, Node's α 0.5 degree, Node's α 1.5 degree), where we use our newly implemented node's measures aggregators (see step 1 of figure 4.1).

3. To quantify how informative a certain network measure is, we pass the generated nodes' measure vectors from the customized Network-Only Link-Based relational classifier to a specified WEKA classifier (either Logistic or J48), where we split the dataset into 10 folds using cross-validation. For example to quantify the informativeness of the *C-degree* measure only, we pass the aggregated nodes' *C-degree* feature vectors to WEKA logistic regression or J48 using the customized Network-Only Link-Based relational classifier (see step 2 and 3 of figure 4.1). Moreover, we can pass combined network measures such as *C-degree* with traditional strength (*C-degree* & S) to classify networked data.

4. We then quantitatively compare the informativeness of two network measures against each other in pairs. For example, we compare the classification accuracies for nodes when using the *C-degree* measure only against using the α 0.5 degree only (see step 4 of figure 4.1). To ensure the reliability of the comparison, we use the statistical significance T-test with cross-validation of 10 folds, where for each of the folds we compute the means difference

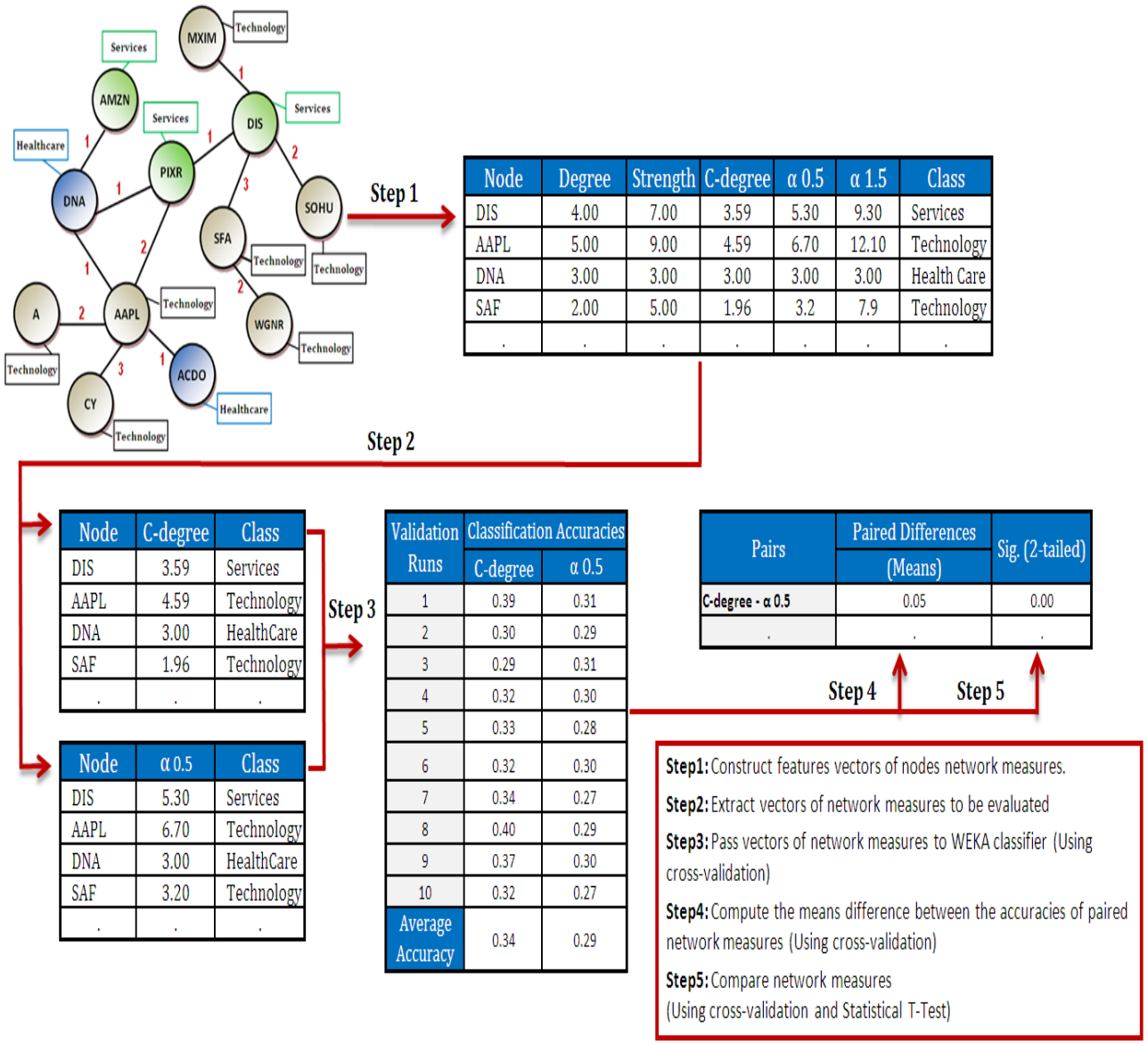


Figure 4.1: Our Comparative Methodology: Example applied to Industry-pr dataset.

between the pairs of accuracies for the paired measures (*C-degree* - α 0.5 degree). Finally, we compute the statistical significance using T-test, in SPSS [1], for the paired network measures on each involved dataset and each used WEKA classifier (see step 5 of figure 4.1).

To evaluate the sensitivity of the collected classification results, in this comparative methodology, to the two used WEKA classifier:

1. We use the same 10 folds cross-validation with statistical significance T-test. For each fold, we compute the means difference between pairs of accuracies generated from logistic regression (LR) and J48 classifiers for each individual network measure on each involved dataset (for example, *C-degree*.J48 - *C-degree*.LR).

2. Finally, we compute the statistical significance using T-tests for paired classifiers for each individual network measure on each involved dataset.

[1] SPSS is Statistical Package for the Social Sciences software available at: http://www.spss.com/

Chapter 5

Analysis and Discussion

In this chapter, we discuss the analysis of our conducted comparative study on two state-of-the-art generalizations of unweighted degree measure on seven benchmark datasets using two WEKA classifiers. To further motivate and to put our results into context, we start in Section 1 by collecting statistics about the involved datasets and exploring their structures. In Section 2, we present the experimental setup. In Section 3, we provide analysis of the results of the comparative methodology on each individual dataset using multiple pairs of network measures and using two WEKA classifiers, logistic regression and J48. We conclude in Section 4 with discussion illustrating how our analysis of the results collected from our comparative methodology answers the research questions we raised in chapter 1.

5.1 Benchmark Datasets Exploration

In this section, we focus on exploring the 7 datasets, before moving to analyze the results of the conducted data classification experiments. We explore the datasets in terms of network structure, with primarily focus on edge weights. To perform this task, we made an intensive use of Microsoft Excel 2007 features such as Data Tools, Data Outline and Functions (Logical, Statistical, Conditional and Lookup functions).

As previously mentioned, we will study 7 labeled datasets from 3 different domains. The first domain is the WebKB with 4 datasets from 3 universities websites (University of Texas, University of Washington and University of Wisconsin), while the second domain is the Industry domain with 2 datasets (Industry-yh and Industry-pr) that represent the relationship between industrial companies from varied industrial sectors that are extracted from news articles. The seventh dataset is from the third domain, the Internet Movie Database (IMDb) website [17].

Data Key	**IMDb**	**Industry**		**WebKB**			
	IMDb-all	Ind.-pr	Ind.-yh	Texas-cocite	Washing.-cocite	Washing.-link	Wiscon.-cocite
Number of Nodes N	1441	2189	1798	338	434	434	354
Number of Class Labels	2	12	12	6	6	6	6
Number of Edges M	51481	13062	14165	32988	30462	1941	33250
% of unweighted 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 5.1: Datasets Statistics

Table 5.1 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 unweighted edges (edges with weight equal one), and the percentage of weighted edges. Table 5.2 illustrates connections in datasets: the average probability that two nodes are connected *connectance*, the percentage of links within sets of nodes of the same class E_{within} , the percentage of links across different classes (communities) of nodes E_{across}, the percentage of weighted links, the percentage of unweighted links, and the percentage of disconnected nodes in each dataset (singletons).

Data Key	IMDb	Industry		WebKB			
	IMDb-all	Ind.-pr	Ind.-yh	Texas-cocite	Washing.-cocite	Washing.-link	Wiscon.-cocite
Connectance	4.96%	0.55%	0.88%	57.92%	32.42%	2.07%	53.22%
% of E_{within}	66.17%	43.15%	40.78%	76.14%	59.63%	27.56%	77.70%
% of $E_{within,unweighted}$	86.91%	67.65%	59.97%	76.08%	59.01%	81.50%	82.52%
% of $E_{within,weighted}$	13.09%	32.35%	40.03%	23.92%	40.99%	18.50%	17.48%
% of E_{across}	33.83%	56.85%	59.22%	23.86%	40.37%	72.44%	22.30%
% of $E_{across,unweighted}$	88.23%	80.11%	67.59%	68.59%	60.55%	89.47%	66.52%
% of $E_{across,weighted}$	11.77%	19.89%	32.41%	31.41%	39.45%	10.53%	33.48%
% of Singletons	4.44%	0.00%	0.00%	1.18%	0.00%	0.23%	1.69%

Table 5.2: Datasets Connections

The WebKB datasets almost show the highest node's connectance, while Industry datasets show the lowest node's connectance over all other datasets. The majority of links in all datasets are unweighted with overall percentage of 59.63% to 87.36% (of which 59.01% to 89.47% are unweighted E_{within} and E_{across}), while both Industry-yh and Washington-cocite have relatively higher variety in weighted connections within and across communities. Whereas, the implementation of the NETKIT-SRL Toolkit focuses on networked data only, all disconnected nodes (singleton) were removed while aggregating features. Our implementation of generalizations aggregators takes singletons into account, therefore, some statistics that we present may slightly differ from those reported in NETKIT-SRL univariate case study [17].

We find it interesting while exploring each dataset to study the correlation between the network structure and the classes of nodes using network measures, as dyadicity Dy and heterophilicity H, and to examine the homophily of data classes. For each dataset in table 5.3, we find the original Dy and H measures [25] along with their generalized versions, $Dy_{\,Generalized}$ and $H_{\,Generalized}$, based on the effective cardinality concept [1] . We try to observe the effect of assigning weights to edges on data correlation by comparing the homophily of each dataset classes when using the generalized and the unweighted graph dyadicity and heterophilicity.

Data Key	IMDb	Industry		WebKB			
	IMDb-all	Ind.-pr	Ind.-yh	Texas-cocite	Washing.-cocite	Washing.-link	Wiscon.-cocite
$Dy_{\,unweighted}$	50%	100%	100%	50%	66.67%	66.67%	50%
$H_{\,unweighted}$	0%	8.33%	8.33%	0%	16.67%	50%	0%
$Homophily_{\,unweighted}$	50%	91.67%	91.67%	50%	50%	33.33%	50%
$Dy_{\,Generalized}$	50%	75.00%	58.33%	33.33%	50%	66.67%	33.33%
$H_{\,Generalized}$	0%	8.33%	0%	0%	0%	33.33%	0%
$Homophily_{\,Generalized}$	50%	66.67%	58.33%	33.33%	50%	33.33%	33.33%

Table 5.3: Dyadicity and Heterophilicity Measures

Table 5.3 illustrates the percentage of classes in each data set that tend to be *dyadic* ($Dy > 1$), *heterophobic* ($H > 1$) and that exhibit Homophily (for both unweighted and generalized cases).

From the rows of unweighted measures $Dy_{\,unweighted}$ and $H_{unweighted}$ in table 5.3 we can observe that, for almost all sectors (91.67% of sectors) in Industry datasets : companies of the same industrial sector exhibit $Homophily_{\,unweighted}$ as they have a clear clustering tendency within their sectors more than the expected for a random configuration, and have fewer connections to companies from other sectors than the expected for random configuration (100% of industrial sectors are dayadic $Dy > 1$ and 91.67% are anti-heterophobic $H < 1$).

The situation differs in the generalized measures case, where $Dy_{\,Generalized}$ and $H_{\,Generalized}$ measures take the weight of links into account and may reflect more accurate correlation between nodes (sensitive to strength of link between nodes). As show in table 5.3 generalized measures leads to decline in number of industrial sectors that exhibit homophily as a result of decline in dyadic classes in both Industry-pr and Industry-yh by 25% and 41.67% respectively (where 32.35% and 40.03% of the E_{within} are weighted in the two datasets respectively, as shown in table 5.2).

From table5.3, IMDb nodes maintains the same clustering tendency in both unweighted and generalized measures (as 86.91% of the E_{within} are unweighted and 88.23% of the E_{across} are unweighted, as shown in table 5.2). The WebKB datasets results show that Washington

datasets maintain the same homophiliy in both unweighted and generalized cases, while the both Texas-cocite and Wisconsin-cocite show decline in the number of classes that exhibit homophiliy in generalized cases by 17%.

As shown in table 5.4, some networked nodes in IMDb and Industry datasets have identical node's traditional unweighted degree and strength (as all edges incident to a node are of weights equal to 1). This finding leads nodes to have the same identical value for all degree generalizations measures that rely on unweighted degree and strength: the *C-degree*, the *α 0.5 degree* and the *α 1.5 degree* (and consequently might have the same classification results when passing different measures as additional features to different classifiers).

Data Key	**IMDb**	**Industry**		**WebKB**		
	IMDb-all	Ind.-pr	Ind.-yh	Texas-cocite	Washington	Wiscon.-cocite
% of $Nodes_{\,Identical\ measures}$	9.16%	48.42%	33.15%	0%	0%	0%

Table 5.4: Nodes of Identical Measures

5.2 Experimental Setup

We conducted several classification experiments using nLB classifier and based on the aggregated network measures vectors: three generalizations (*C-degree*, α 0.5 degree, α 1.5 degree), two traditional network measures (Node's degree and Node's strength), and two combinations of network measures (*C-degree* with strength $CD\&S$ and Degree with strength $D\&S$).

We split each dataset using cross-validation into 10 folds, and then we run classification experiments based on each network measure alone using the customized Network-Only Link-Based Classifier with each of the tested WEKA classifiers (logistic regression and J48). The final accuracy for each experiment is averaged over 10 for each measure as shown in logistic regression results in Appendix B and in J48 results in Appendix C. We used the improvement in the classification accuracy as an indicator of the efficiency of the additional information provided by each network measure.

To compare quantitatively the informativeness of different degree generalizations with the original unweighted degree measure and against other generalizations, and to increase the reliability of our results, we split each dataset into 10 folds using cross-validation, then we compute the means difference between pairs of network measures for each fold. Finally, we compute the statistical significance T-tests for each pair using one WEKA classifier only at a time. To evaluate the sensitivity of the collected classification results in this comparative methodology to the used WEKA classifiers, we compute the statistical significance T-tests for pairs of classification results based on each individual network measure using two WEKA classifiers on each involved dataset.

5.3 Datasets Analysis

In this section, we present the analysis of the results of our comparative methodology to compare quantitatively the informativness of each of the two-state-of-the-art generalizations of unweighted degree measure using within-network classification on each involved dataset separately, using logistic regression and J48 classifiers. For each involved dataset, we present the results of comparing quantitatively network measures using the means difference of classification accuracies and the statistical significance T-test of the paired measures. Moreover, we evaluate the efficiency of combining more than one network measure as additional information for data classification.

5.3.1 Industry Datasets

Industry-pr Dataset

The results of the 21 pairs significance T-test that we apply on logistic regression classification experiments' results for Industry-pr are recorded as pairs of network measures in table 5.5, while the J48 significance T-test results are recorded in table 5.6.

Pairs		Paired Difference (Mean)	Significance Test (2-tailed)
Pair 1	D - CD	-6.442	.000
Pair 2	D - S	-2.22	.000
Pair 3	D - $CD\&S$	-8.72	.000
Pair 4	D - $D\&S$	-2.15	.000
Pair 5	CD - S	4.25	.006
Pair 6	CD - $CD\&S$	-.2.23	.133
Pair 7	CD - $D\&S$	4.35	.002
Pair 8	S - $CD\&S$	-6.53	.000
Pair 9	S - $D\&S$	0.05	.921
Pair 10	$CD\&S$ -$D\&S$	6.67	.000
Pair 11	α 0.5 - α 1.5	-0.55	.309
Pair 12	α 0.5 - D	1.81	.000
Pair 13	α 0.5 - S	-0.41	.316
Pair 14	α 0.5 - CD	-4.66	.003
Pair 15	α 0.5 - $D\&S$	-0.36	.443
Pair 16	α 0.5 - $CD\&S$	-6.94	.000
Pair 17	α 1.5 - D	2.33	.000
Pair 18	α 1.5 - S	0.14	.722
Pair 19	α 1.5 - CD	-4.11	.007
Pair 20	α 1.5 - $D\&S$	0.22	.686
Pair 21	α 1.5 - $CD\&S$	-6.45	.000

Table 5.5: Logistic Regression Significance T-test Results: Industry-pr

Pairs		Paired Difference (Mean)	Significance Test (2-tailed)
Pair 1	D - CD	-10.23	.000
Pair 2	D - S	0.14	0.856
Pair 3	D - $CD\&S$	-11.79	.000
Pair 4	D - $D\&S$	0.64	0.539
Pair 5	D - α 0.5	1.65	0.117
Pair 6	D - α 1.5	0.55	0.576
Pair 7	CD - S	10.37	.000
Pair 8	CD - $CD\&S$	-1.55	0.38
Pair 9	CD - $D\&S$	10.87	.000
Pair 10	CD - α 0.5	11.88	.000
Pair 11	CD - α 1.5	10.78	.000
Pair 12	S - $CD\&S$	-11.92	.000
Pair 13	S - $D\&S$	0.50	0.625
Pair 14	S - α 0.5	1.51	0.12
Pair 15	S - α 1.5	0.41	0.512
Pair 16	$CD\&S$ - $D\&S$	12.43	.000
Pair 17	$CD\&S$ - α 0.5	13.43	0.000
Pair 18	$CD\&S$ - α 1.5	12.34	.000
Pair 19	$D\&S$ - α 0.5	1.01	0.334
Pair 20	$D\&S$ - α 1.5	-0.09	0.946
Pair 21	α 0.5 - α 1.5	-1.10	0.303

Table 5.6: J48 Significance T-test Results: Industry-pr

From Industry-pr T-tests tables 5.5 and 5.6, we can see that classification accuracies based on node's continuous degree *C-degree* significantly outperforms almost all classification accuracies based on the other network measures with significances varies from 0.0% to 0.7%, except the classifications based on the combination of node's *C-degree* and strength $CD\&S$ using the two classifiers.

From tables 5.5 and 5.6, we can see that combining node's continuous degree *C-degree* with node's strength $CD\&S$ shows significant improvement in classification accuracies over all other network measures with high significance of 0.0% for both logistic regression and J48 classifiers,

except the *C-degree* disjointedly (as $CD\&S$ outperforms *C-degree* without significance). Moreover, $CD\&S$ shows on average the highest accuracies in experimental runs for both logistic and J48 classifiers as shown in classification results tables C.6 and B.6.

It is worth noting that when using logistic regression the node's degree D based classification is outperformed by all other generalized network measures with a very high significance of 0.0%, while the node's strength S never show any significant improvement over any generalizations. However, combining node's strength with node's degree in the $D\&S$, the α 0.5 degree and the α 1.5 degree shows significant improvement over the original unweighted degree D, while the generalized α 1.5 degree shows significant improvement over the S too. These results emphasis on complementing unweighted degree measure or generalized degree measures with node's strength for better patterns discovery.

Industry-yh Dataset

Pairs		Paired Difference (Mean)	Significance Test (2-tailed)
Pair 1	D - CD	-7.73	.001
Pair 2	D - S	-0.28	.712
Pair 3	D - $CD\&S$	-7.90	.000
Pair 4	D - $D\&S$	-0.28	.712
Pair 5	CD - S	7.40	.000
Pair 6	CD - $CD\&S$	-0.22	.852
Pair 7	CD - $D\&S$	7.40	.000
Pair 8	S - $CD\&S$	-7.63	.000
Pair 10	$CD\&S$ -$D\&S$	7.63	.000
Pair 12	α 0.5 - D	0.28	.712
Pair 13	α 0.5 - CD	-7.40	.000
Pair 15	α 0.5 - $CD\&S$	-7.63	.000
Pair 17	α 1.5 - D	0.28	.712
Pair 18	α 1.5 - CD	-7.40	.000
Pair 20	α 1.5 - $CD\&S$	-7.63	.000

Table 5.7: Logistic Regression Significance T-test Results: Industry-yh

Pairs		Paired Difference (Mean)	Significance Test (2-tailed)
Pair 1	D - CD	-2.51	0.089
Pair 2	D - S	2.61	0.041
Pair 3	D - $CD\&S$	-4.40	0.011
Pair 4	D - $D\&S$	1.73	0.244
Pair 5	D - α 0.5	2.56	0.022
Pair 6	D - α 1.5	3.23	0.009
Pair 7	CD - S	5.12	0.002
Pair 8	CD - $CD\&S$	-1.89	0.256
Pair 9	CD - $D\&S$	4.23	0.005
Pair 10	CD - α 0.5	5.06	0.002
Pair 11	CD - α 1.5	5.74	0.003
Pair 12	S - $CD\&S$	-7.01	.000
Pair 13	S - $D\&S$	-0.89	0.489
Pair 14	S - α 0.5	-0.06	0.939
Pair 15	S - α 1.5	0.61	0.586
Pair 16	$CD\&S$ - $D\&S$	6.13	0.002
Pair 17	$CD\&S$ - α 0.5	6.95	.000
Pair 18	$CD\&S$ - α 1.5	7.63	.000
Pair 19	$D\&S$ - α 0.5	0.83	0.575
Pair 20	$D\&S$ - α 1.5	1.50	0.215
Pair 21	α 0.5 - α 1.5	0.67	0.43

Table 5.8: J48 Significance T-test Results: Industry-yh

From Industry-yh T-tests tables 5.7 and 5.8, we can see that classification accuracies based on the node's continuous degree *C-degree* in both logistic regression and J48 significantly outperforms the node's strength, the α 0.5 degree, the α 1.5 degree and the combination of node's degree and strength $D\&S$ (with high significance of 0.0% when using logistic regression classifier and with significance between 0.0% and 0.3% when using J48). Moreover the node's

C-degree significantly outperforms the original unweighted degree D when using the logistic regression classifier.

As happened in Industry-pr, it is clear that combining node's *C-degree* with node's strength in $CD\&S$ shows no significant improvement over the *C-degree* based classifications, while the $CD\&S$ significantly outperforms all other classification accuracies based on other network measures with high significance of 0.0% when using logistic regression classifier and with significance between 0.0% and 0.2% when using J48. Both classification accuracies based on node's *C-degree* and based on the combination of node's strength and *C-degree* outperforms the node's degree based classification accuracy with high significance of 0.0% as in table 5.7. Moreover, the node's degree shows on average the worst classification accuracies over all other network measures (see the results of experimental runs using logistic regression in table B.7)

From logistic regression T-tests table 5.7, it is worth noting that six pairs are missing from significance T-test results: pair 9 (node's strength *vs.* degree with strength), pair 11 (node's α 0.5 degree *vs.* α 1.5 degree), pair 14 (node's α 0.5 degree *vs.* strength), pair 16 (node's α 0.5 *vs.* degree with strength), pair 19 (node's α 1.5 *vs.* degree with strength), and pair 21 (node's α 1.5 *vs.* degree with strength). The reason behind that is the equal classification accuracies results for these network measures (see the results of experimental runs table B.7). This finding can emphasis on combining node's degree and strength in generalized continuous degree *C-degree* captures the focus of interaction between nodes much better than alpha degree generalizations and combination of the original degree with strength $D\&S$, where *C-degree* outperforms all generalized network measures.

5.3.2 WebKB Datasets

Texas-cocite Dataset

Pairs		Paired Difference (Mean)	Significance Test (2-tailed)
Pair 1	D - CD	4.76	.026
Pair 2	D - S	3.54	.081
Pair 3	D - $CD\&S$	-1.53	.484
Pair 4	D - $D\&S$	3.55	.0294
Pair 5	CD - S	-1.22	.588
Pair 6	CD - $CD\&S$	-6.32	.066
Pair 7	CD - $D\&S$	-1.21	.5380
Pair 8	S - $CD\&S$	-5.07	.063
Pair 9	S - $D\&S$	0.01	.9966
Pair 10	$CD\&S$ - $D\&S$	5.08	.0666
Pair 11	α 0.5 - D	-0.60	.744
Pair 12	α 0.5 - S	2.00	.175
Pair 13	α 0.5 - CD	4.29	.096
Pair 14	α 0.5 - $CD\&S$	-2.10	.337
Pair 15	α 0.5 - $D\&S$	.2.00	.0723
Pair 16	α 0.5 - α 1.5	5.64	.040
Pair 17	α 1.5 - D	-6.21	.019
Pair 18	α 1.5 - S	-2.74	.225
Pair 19	α 1.5 - CD	-1.53	.666
Pair 20	α 1.5 - $CD\&S$	-7.75	.037
Pair 21	α 1.5 - $D\&S$	-2.66	.2022

Table 5.9: Logistic Regression Significance T-test Results: Texas-cocite

From J48 T-tests table 5.10, we can see that the *C-degree* based classification accuracies significantly outperform both the node's strength and the α 1.5 degree, while it never outperform any other measure significantly when using logistic regression.

On the one hand, the classification accuracies based on the combination of node's strength and node's *C-degree* ($CD\&S$) significantly outperform the classification accuracies based on node's α 1.5 degree in both logistic regression and J48 with significance of 3.7% and 0.1% respectively, while it outperform node's strength based accuracies with significance of 0.3% using J48 classifier. Moreover the $CD\&S$ shows improvement (without significance) on average over

Pairs		Paired Difference (Mean)	Significance Test (2-tailed)
Pair 1	D - CD	-0.20	0.935
Pair 2	D - S	10.40	0.006
Pair 3	D - $CD\&S$	-0.80	0.696
Pair 4	D - $D\&S$	-2.60	0.462
Pair 5	D - α 0.5	3.60	0.294
Pair 6	D - α 1.5	13.40	0.004
Pair 7	CD - S	10.60	0.001
Pair 8	CD - $CD\&S$	-0.60	0.808
Pair 9	CD - $D\&S$	-2.40	0.447
Pair 10	CD - α 0.5	3.80	0.162
Pair 11	CD - α 1.5	13.60	.000
Pair 12	S - $CD\&S$	-11.20	0.003
Pair 13	S - $D\&S$	-13.0	0.005
Pair 14	S - α 0.5	-6.80	0.01
Pair 15	S - α 1.5	3.00	0.331
Pair 16	$CD\&S$ - $D\&S$	-1.80	0.572
Pair 17	$CD\&S$ - α 0.5	4.40	0.176
Pair 18	$CD\&S$ - α 1.5	14.20	0.001
Pair 19	$D\&S$ - α 0.5	6.20	0.047
Pair 20	$D\&S$ - α 1.5	16.00	0.002
Pair 21	α 0.5 - α 1.5	9.80	0.009

Table 5.10: J48 Significance T-test Results: WebKB-Texas-cocite

each of the *C-degree* and strength disjointedly in both logistic and J48 classification experimental runs as shown in tables B.1 and C.1.

On the other, combining node's original degree with strength in $D\&S$ improves the classification based on the node's strength to outperform the node's α 0.5 degree, the node's α 1.5 degree and the node's strength using the J48 classifier, where the strength disjointedly never outperform any other measure using any classifier, while the node's degree outperforms the α 1.5 degree using both logistic and J48 and outperforms strength using J48.

The node's α 1.5 degree based classification shows on average the worst classification accuracies over all other network measures in both logistic and J48 classification experiments (see the results of experimental runs in tables B.1 and C.1). Moreover, as shown in T-tests tables 5.10 and 5.9, the node's degree and the α 0.5 degree outperform the α 1.5 degree with significance in both logistic regression and J48 cases and the *C-degree* outperforms the α 1.5 degree in logistic regression only.

Washington-cocite Dataset

As clear from J48 T-tests table 5.12 , Washington-cocite dataset never show any significant results for all T-tests on different network measures using Decision Trees. While when using logistic regression, the nod's *C-degree* is outperformed by all other network measures with significances varying from 0.0% to 1.6% as shown in table 5.11. Moreover, from the results of logistic regression experimental runs in table B.2, we can see that node's continuous degree *C-degree* based classification shows the worst classification accuracies on average over all other accuracies.

In logistic regression classifications, the node's degree based accuracies significantly outperforms the nodes' *C-degree* with 0.6%, while combining node's strength with node's degree measure shows significant improvement of 0.2% in classification accuracies over the node's *C-degree* and over the α 1.5 degree with 3.8 %.

Washington-Link1 Dataset

From tables 5.13 and 5.14, we can see that node's continuous degree *C-degree* based classification significantly outperforms classifications based on the node's strength, the α 0.5 degree and the α 1.5 degree using both logistic regression and J48, while the *C-degree* also significantly outperforms the node's original degree in logistic regression only. Both node's original degree

Pairs		Paired Difference (Mean)	Significance Test (2-tailed)
Pair 1	D - CD	9.01	.006
Pair 2	D - S	-0.45	.839
Pair 3	D - $CD\&S$	0.47	.834
Pair 4	D - $D\&S$	-0.42	.887
Pair 5	CD - S	-9.46	.001
Pair 6	CD - $CD\&S$	-8.55	.016
Pair 7	CD - $D\&S$	-9.43	.002
Pair 8	S - $CD\&S$	0.91	.645
Pair 9	S - $D\&S$	0.03	.987
Pair 10	$CD\&S$ - $D\&S$	-0.88	.738
Pair 11	α 0.5 - α 1.5	3.45	.083
Pair 12	α 0.5 - D	-0.51	.839
Pair 13	α 0.5 - S	-0.94	.568
Pair 14	α 0.5 - CD	8.52	.000
Pair 15	α 0.5 - $CD\&S$	-0.03	.992
Pair 16	α 0.5 - $D\&S$	-0.91	.718
Pair 17	α 1.5 - D	-3.94	.140
Pair 18	α 1.5 - S	-4.47	.020
Pair 19	α 1.5 - CD	5.07	.009
Pair 20	α 1.5 - $CD\&S$	-3.53	.161
Pair 21	α 1.5 - $D\&S$	-4.40	.038

Table 5.11: Logistic Regression Significance T-test Results: Washington-cocite

Pairs		Paired Difference (Mean)	Significance Test (2-tailed)
Pair 1	D - CD	2.12	0.448
Pair 2	D - S	3.43	0.118
Pair 3	D - $CD\&S$	4.12	0.153
Pair 4	D - $D\&S$	1.131	0.747
Pair 5	D - α 0.5	0.43	0.834
Pair 6	D - α 1.5	3.57	0.238
Pair 7	CD - S	1.31	0.696
Pair 8	CD - $CD\&S$	2.00	0.596
Pair 9	CD - $D\&S$	-0.99	0.807
Pair 10	CD - α 0.5	-1.71	0.649
Pair 11	CD - α 1.5	1.34	0.649
Pair 12	S - $CD\&S$	0.72	0.743
Pair 13	S - $D\&S$	-2.3	0.354
Pair 14	S - α 0.5	-3.00	0.109
Pair 15	S - α 1.5	0.026	0.99
Pair 16	$CD\&S$ - $D\&S$	-2.992	0.387
Pair 17	$CD\&S$ - α 0.5	-3.75	0.134
Pair 18	$CD\&S$ - α 1.5	-0.66	0.762
Pair 19	$D\&S$ - α 0.5	-0.70	0.775
Pair 20	$D\&S$ - α 1.5	2.33	0.339
Pair 21	α 0.5 - α 1.5	3.03	0.166

Table 5.12: J48 Significance T-test Results: WebKB-Washington-cocite

and node's strength shows no any significant improvement in classification accuracies over generalized network measures using both logistic regression and J48 classifiers.

On the one hand, combining node's strength with continuous degree *C-degree* $CD\&S$, shows significant improvement over the original degree, the α 0.5 degree and the α 1.5 degree in both logistic regression and J48 classifications, while the $CD\&S$ also significantly outperforms the node's strength in logistic regression classification only. Moreover in J48, the classification accuracies based on $CD\&S$ significantly outperforms the classification accuracies based on the combination of node's original degree and strength $D\&S$ with significance of 3.6%.

On the other hand, combining node's strength with node's degree $D\&S$, shows significant improvement over the node's degree, the strength, the α 0.5 degree and the α 1.5 degree in logistic regression classifications, while it shows no any significant improvement over other measures in case of J48 classification.

Pairs		Paired Difference (Mean)	Significance Test (2-tailed)
Pair 1	D - CD	-5.35	.149
Pair 2	D - S	2.79	.406
Pair 3	D - $CD\&S$	-6.01	.029
Pair 4	D - $D\&S$	-13.39	.0050
Pair 5	CD - S	7.94	.008
Pair 6	CD - $CD\&S$	-0.73	.822
Pair 7	CD - $D\&S$	-7.98	.0585
Pair 8	S - $CD\&S$	-8.64	.015
Pair 9	S - $D\&S$	-15.84	.0057
Pair 10	$CD\&S$ - $D\&S$	-7.35	.0590
Pair 11	α 0.5 - D	-2.85	.182
Pair 12	α 0.5 - S	-0.21	.945
Pair 13	α 0.5 - CD	-8.10	.007
Pair 14	α 0.5 - $D\&S$	-16.14	.0004
Pair 15	α 0.5 - $CD\&S$	-8.89	.013
Pair 16	α 0.5 - α 1.5	1.28	.607
Pair 17	α 1.5 - D	-3.93	.186
Pair 18	α 1.5 - S	-1.36	.748
Pair 19	α 1.5 - CD	-9.22	.030
Pair 20	α 1.5 - $CD\&S$	-9.95	.024
Pair 21	α 1.5 - $D\&S$	-17.20	.0009

Table 5.13: Logistic Regression Significance T-test Results: Washington-link

Pairs		Paired Difference (Mean)	Significance Test (2-tailed)
Pair 1	D - CD	-12.43	0.001
Pair 2	D - S	-0.95	0.769
Pair 3	D - $CD\&S$	-9.44	0.005
Pair 4	D - $D\&S$	-1.09	0.715
Pair 5	D - α 0.5	0.01	0.996
Pair 6	D - α 1.5	-0.01	0.997
Pair 7	CD - S	11.48	0.01
Pair 8	CD - $CD\&S$	2.99	0.363
Pair 9	CD - $D\&S$	11.34	0
Pair 10	CD - α 0.5	12.44	0.002
Pair 11	CD - α 1.5	12.42	0.004
Pair 12	S - $CD\&S$	-8.49	0.07
Pair 13	S - $D\&S$	-0.14	0.976
Pair 14	S - α 0.5	0.96	0.749
Pair 15	S - α 1.5	0.94	0.694
Pair 16	$CD\&S$ - $D\&S$	8.35	0.036
Pair 17	$CD\&S$ - α 0.5	9.45	0.019
Pair 18	$CD\&S$ - α 1.5	9.43	0.008
Pair 19	$D\&S$ - α 0.5	1.10	0.719
Pair 20	$D\&S$ - α 1.5	1.08	0.804
Pair 21	α 0.5 - α 1.5	-0.02	0.995

Table 5.14: J48 Significance T-test Results: WebKB-Washington-link

Wisconsin-cocite Dataset

From tables 5.15 and 5.16,we can see that node's continuous degree *C-degree* shows no any significant improvement over any other network measures for both logistic regression and J48, while combining node's strength with node's *C-degree* leads to significant improvement over the α 1.5 degree using both logistic regression and J48.

The classification accuracies based on the node's degree significantly outperforms the classification accuracies based on the node's continuous degree *C-degree* and based on the α 1.5 degree with high significances ranging from 0.1% and 0.4% for both logistic and J48 classifiers, whereas combining node's strength with node's degree $D\&S$, shows significant improvement over the node's *C-degree*, the strength and the α 1.5 degree in J48 classification, while it shows no any significant improvement over other measures in case of logistic regression classification.

Pairs		Paired Difference (Mean)	Significance Test (2-tailed)
Pair 1	D - CD	6.98	.002
Pair 2	D - S	3.47	.090
Pair 3	D - $CD\&S$	2.02	.572
Pair 4	D - $D\&S$	3.65	.0977
Pair 5	CD - S	-3.51	.099
Pair 6	CD - $CD\&S$	-4.92	.158
Pair 7	$CD\&S$ - $D\&S$	1.63	.6419
Pair 8	S - $CD\&S$	-1.41	.578
Pair 9	S - $D\&S$	0.34	.9107
Pair 10	CD - $D\&S$	-3.22	.2337
Pair 11	α 0.5 - D	-1.41	.245
Pair 12	α 0.5 - CD	.5.56	.009
Pair 13	α 0.5 - S	1.98	.189
Pair 14	α 0.5 - $CD\&S$	0.60	.830
Pair 15	α 0.5 - $D\&S$	2.24	.2651
Pair 16	α 0.5 - α 1.5	7.97	.002
Pair 17	α 1.5 - D	-9.49	.001
Pair 18	α 1.5 - CD	-2.52	.225
Pair 19	α 1.5 - S	-5.99	.011
Pair 20	α 1.5 - $CD\&S$	-7.43	.009
Pair 21	α 1.5 - $D\&S$	-5.74	.0524

Table 5.15: Logistic Regression Significance T-test Results: Wisconsin-cocite

Pairs		Paired Difference (Mean)	Significance Test (2-tailed)
Pair 1	D - CD	4.84	0.041
Pair 2	D - S	5.95	0.016
Pair 3	D - $CD\&S$	2.66	0.24
Pair 4	D - $D\&S$	1.13	0.66
Pair 5	D - α 0.5	3.79	0.212
Pair 6	D - α 1.5	5.41	0.014
Pair 7	CD - S	1.11	0.666
Pair 8	CD - $CD\&S$	-2.36	0.182
Pair 9	CD - $D\&S$	-3.71	0.081
Pair 10	CD - α 0.5	-1.23	0.676
Pair 11	CD - α 1.5	0.54	0.756
Pair 12	S - $CD\&S$	-3.47	0.098
Pair 13	S - $D\&S$	-4.83	0.025
Pair 14	S - α 0.5	-02.34	0.367
Pair 15	S - α 1.5	-0.61	0.735
Pair 16	$CD\&S$ - $D\&S$	-1.43	0.36
Pair 17	$CD\&S$ - α 0.5	1.10	0.553
Pair 18	$CD\&S$ - α 1.5	2.83	0.053
Pair 19	$D\&S$ - α 0.5	2.53	0.175
Pair 20	$D\&S$ - α 1.5	4.34	0.009
Pair 21	α 0.5 - α 1.5	1.72	0.476

Table 5.16: J48 Significance T-test Results: WebKB-Wisconsin-cocite

5.3.3 IMDb-all Dataset

IMDb-all dataset shows no any significant results for all T-tests on all network measures using both logistic regression and J48 classifiers (see tables 5.17 and 5.18). This may be due to the nature of the IMDb network, the way of creating links between movies in this datasets or the small space for improvement in accuracies due to high base accuracy (0.66) and high classification accuracies for this dataset (ranging from 60% up to 78.47%) as will be discussed in section 5.4.

From table 5.17, it is worth noting that pair 6 (node's strength *vs.* node's degree with strength) is missing from logistic regression significance test, as combining node's strength and node's original degree in ($D\&S$) leads on average to equal classification accuracy results (see the results of experimental runs in tables B.5 and C.5).

Pairs		Paired Difference (Mean)	Significance Test (2-tailed)
Pair 1	D - CD	0.63	.739
Pair 2	D - S	2.01	.343
Pair 3	D - $CD\&S$	0.63	.739
Pair 4	D - $D\&S$	1.81	.262
Pair 5	CD - S	1.34	.495
Pair 7	CD - $D\&S$	1.22	.594
Pair 8	S - $CD\&S$	-1.38	.495
Pair 9	S - $D\&S$	-0.21	.843
Pair 10	$CD\&S$ - $D\&S$	1.18	.594
Pair 11	α 0.5 - D	-0.38	.876
Pair 12	α 0.5 - CD	0.35	.874
Pair 13	α 0.5 - S	1.73	.511
Pair 14	α 0.5 - $CD\&S$	0.35	.874
Pair 15	α 0.5 - $D\&S$	1.53	.474
Pair 16	α 0.5 - α 1.5	3.10	.244
Pair 17	α 1.5 - D	-3.33	.124
Pair 18	α 1.5 - CD	-2.71	.239
Pair 19	α 1.5 - S	-1.32	.295
Pair 20	α 1.5 - $CD\&S$	-2.71	.239
Pair 21	α 1.5 - $D\&S$	-1.53	.248

Table 5.17: Logistic Regression Significance T-test Results: IMDb-all

Pairs		Paired Difference (Mean)	Significance Test (2-tailed)
Pair 1	D - CD	0.76	0.691
Pair 2	D - S	-0.69	0.751
Pair 3	D - $CD\&S$	1.67	0.431
Pair 4	D - $D\&S$	-0.97	0.632
Pair 5	D - α 0.5	-1.40	0.362
Pair 6	D - α 1.5	-0.49	0.829
Pair 7	CD - S	-1.46	0.466
Pair 8	CD - $CD\&S$	0.90	0.655
Pair 9	CD - $D\&S$	-1.74	0.251
Pair 10	CD - α 0.5	-2.23	0.355
Pair 11	CD - α 1.5	-1.25	0.479
Pair 12	S - $CD\&S$	2.41	0.281
Pair 13	S - $D\&S$	-0.30	0.783
Pair 14	S - α 0.5	-0.74	0.697
Pair 15	S - α 1.5	0.21	0.93
Pair 16	$CD\&S$ - $D\&S$	-2.64	0.254
Pair 17	$CD\&S$ - α 0.5	-3.16	0.072
Pair 18	$CD\&S$ - α 1.5	-2.23	0.356
Pair 19	$D\&S$ - α 0.5	-0.42	0.823
Pair 20	$D\&S$ - α 1.5	0.48	0.782
Pair 21	α 0.5 - α 1.5	0.90	0.663

Table 5.18: J48 Significance T-test Results: IMDb

5.4 Discussion of Research Questions

In this section, we discuss how our analysis of the results collected from our comparative methodology answers the research questions we raised in Chapter 1.

Table 5.19 summarizes the average difference (mean) in classification accuracies between pairs of network measures on 7 datasets using the logistic regression classifier, while table 5.20 shows similar differences but using the decision tree (J48) classifier. A dash appears in a table cell refers to insignificant average difference between a pair of network measures for a particular dataset over 10 folds, while a numerical value in a table cell represents the significant average difference between a pair of network measures.

Given a generalization of an unweighted network measure, does the generalization provide more information than the original unweighted measure?

Referring to logistic regression results in table 5.19, we notice that neither the *C-degree* nor the α*-degree* results show consistent higher accuracy against the unweighted degree measure, where we use the improvement in the classification accuracy as an indication of the additional

Measures Pairs	IMDb	Industry		WebKB			
	IMDb-all	Ind.-pr	Ind.-yh	Texas-cocite	Washing.-cocite	Washing.-link	Wiscon.-cocite
D - S	-	-2.2%	-	-	-	-	-
CD - D	-	6.4%	7.7%	-4.8%	-9.0%	-	-6.9%
α 0.5 - D	-	1.8%	-	-	-	-	-
α 1.5 - D	-	2.3%	-	-6.2%	-	-	-9.4%
CD - S	-	4.2%	7.4%	-	-9.5%	7.9%	-
α 0.5 - S	-	-	-	-	-	-	-
α 1.5 - S	-	-	-	-	-4.4%	-	-6.6%
CD - α 0.5	-	4.7%	7.4%	-	-8.5%	8.1%	-5.5%
CD - α 1.5	-	-	7.4%	-	-5.1%	9.2%	-
α 0.5 - α 1.5	-	-	-	5.6%	-	-	-
CD&S -D	-	8.7%	7.9%	-	-	6.0%	-
CD&S -S	-	6.5%	7.6%	-	-	8.6%	-
CD&S -D&S	-	6.6%	7.6%	-	-	-	-
CD&S - α 0.5	-	6.9%	7.6%	-	-	8.8%	-
CD&S - α 1.5	-	6.4%	7.6%	7.7%	-	9.9%	7.4%

Table 5.19: 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	-	-	2.6%	10.4%	-	-	6.0%
CD - D	-	10.2%	-	-	-	12.4%	-4.8%
α 0.5 - D	-	-	-2.6%	-	-	-	-
α 1.5 - D	-	-	-3.2%	-13.4%	-	-	-5.4%
CD - S	-	10.4%	5.1%	10.6%	-	11.5%	-
α 0.5 - S	-	-	-	-6.8%	-	-	-
α 1.5 - S	-	-	-	-	-	-	-
CD - α 0.5	-	5.1%	11.9%	-	-	12.4%	-
CD - α 1.5	-	5.7%	10.8%	13.6%	-	12.4%	-
α 0.5 - α 1.5	-	-	-	9.8%	-	-	-
CD&S -D	-	11.8%	4.4%	-	-	9.4%	-
CD&S -S	-	11.9%	7.0%	11.2%	-	8.5%	-
CD&S -D&S	-	12.4%	6.1%	-	-	8.4%	-
CD&S - α 0.5	-	7.0%	13.4%	-	-	9.5%	-
CD&S - α 1.5	-	7.6%	12.3%	14.2%	-	9.4%	-

Table 5.20: J48: Significance Test Results

information provided by the generalization. This can be clearly noticed from the mixed negative and positive differences shown in the pairs: $CD - D$, α 0.5 $- D$ and α 1.5 $- D$. The J48 classifier in table 5.20 shows similar results, but with great advantage to 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 both 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 which depends mainly on the tuning parameter α to mix both the unweighted degree and the strength, unlike the *C-degree* generalization which depends on the value of each edge weight. Moreover, we observe that the unweighted degree can still be more informative than generalized network measures, based on the type of dataset involved and the used classifier. This observation explains the reason behind the common use of unweighted degree in analyzing networks.

To link between classes homophily in table 5.3 and the classification results, we find that in WebKB datasets, both Texas and Wisconsin experience a drop in the homophily when moving from the unweighted to the generalized measures based on effective cardinality. Accordingly, the node's unweighted degree shows significant performance over the generalized degree *C-degree* using logistic regression for both Texas and Wisconsin datasets and using the J48 as well for Wisconsin only.

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?

On the one hand, logistic regression results in table 5.19 show that there is no consistent

performance for any of the two generalizations, the *C-degree* and the α-*degree*, over all datasets using logistic regression. This can be seen again from mixed negative and positive differences in the pairs: $CD - \alpha\ 0.5$, $CD - \alpha\ 1.5$. On the other, J48 classifier results in table 5.20 show clearer advantage to the *C-degree* as it outperforms the α-*degree* (for both $\alpha = 0.5$ and $\alpha = 1.5$).

Overall, we notice that the *C-degree* outperforms the α-*degree*, if we choose the best classifier. From logistic regression and J48 tables we can tell that the *C-degree* generalization shows better performance than the α-*degree* generalization, as it outperforms the α 0.5 degree in the two Industry datasets (Industry-pr and Industry-yh) and Washington-link using both logistic regression and J48 compared to α 0.5 degree outperforms the *C-degree* in Washington-cocite and Wisconsin-cocite only using logistic regression. Moreover, the *C-degree* generalization outperforms the α 1.5 in four datasets compared to α 1.5 outperforms the *C-degree* in Washington-cocite dataset only.

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.

It is worth noting that the classification accuracies when passing both the *C-degree* and the strength as node features (*C-degree* complemented with strength in $CD\&S$) show consistent higher performance than all examined measures: α-*degree* , the traditional unweighted degree, the strength and even the complemented unweighted degree with strength $D\&S$ (see the corresponding rows at the end of both logistic regression and J48 tables)

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

Results in tables 5.19 and 5.20, emphasize that the type of the dataset involved affects the effectiveness of generalizations. For example, the *C-degree* provides more information than the traditional strength and the α 0.5 degree in the two Industry datasets and Washington-link dataset using logistic and J48, while other WebKB datasets show inconsistent *C-degree* results. 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 doesn't 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.

Overall, the Industry domain 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 has no effect on the effectiveness of different network measures as it maintains the same statistically insignificant results for both logistic regression and J48 classifier. Adding to that, no any clear link between the behavior (results) of any dataset and its statistics (gathered in tables 5.2 and 5.1), specially for the percentages of weighted and unweighted links in each dataset.

Notably, three main factors may affect the dataset behavior: the way in which the dataset's nodes are connected (type of relationship between nodes), the method of assigning weights to connections between nodes (relation strength and interaction focus) and the base accuracy for the involved dataset (base accuracy controls the available space for improvement in classification accuracies while passing different features).

The inconsistency in WebKB datasets behavior reflects the insignificant effect of edge weights. This may be due to the method of assigning weights to co-citation between web pages (link weight has no clear representation of relationship between 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 [17]. In contrast, the industrial domain shows more consistency in results compared to WebKB. Where in companies network, 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 [17]. Furthermore as seen from table 5.3, companies in Industry dataset have the clearest clustering tendency in both cases of unweighted and generalized features. Moreover, the low base accuracy for Industry datasets (around 28%) lead to wider space for improvement when passing different measures to different classifiers.

After we study the effect of the type of dataset (the network) involved on the efficiency of generalizations, it is worth noting to study the effect of the type of the used classifiers on generated results. Table 5.21 below summarizes the differences in the average classification accuracy between pairs of results using J48 classifier and logistic regression classifier (LR) for each individual network measure on each individual involved dataset (only statistically significant differences are shown).

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 5.21: J48-logistic regression: Significance Test Results

From results in table 5.21, we conclude that the performance of any classifier depends on the dataset involved rather than the specific network measures. For example, in all WebKB cocite networks (Texas, Washington, Wisconsin), the J48 classifier adds advantage to the *C-degree* performance over the logistic regression classifier. The IMDb dataset shows insignificant differences between J48 and logistic regression results for all examined pairs of network measures, where it maintains to show statistically insignificant results for all paired measures using either J48 or logistic regression (see tables 5.19 and 5.20). Overall, it is clear that the decision tree classifier J48 outperforms the logistic regression in all examined network measures, except for the Industry-yh and the Washington-link.

Chapter 6

Conclusion

Motivated by the fact that generalizations comparison up until now relied primarily on visual inspection of different plots and informal articulation on how a particular generalization is more informative than the original unweighted measure, in this thesis we have provided an automated comparative methodology for comparing quantitatively two state-of-the-art generalizations of unweighted degree measures, the *C-degree* and the *α-degree*. We have surveyed most recent researchers' attempts to generalize unweighted network measures. We have developed new software components to compare between these generalizations based on classification accuracies, the main indicator of the amount of the additional information provided generalizations. We then conducted an extensive experimental evaluation of the effectiveness of the *C-degree* and the *α-degree* generalizations, when quantitatively compared to each other and against the traditional unweighted degree and strength measures. Also, we have a detailed evaluation of the effect of the type of the dataset involved and the used classifiers on the effectiveness of generalizations. We have utilized a very useful modular network learning toolkit (NetKit-SRL) to get the advantages of the use of common platform which enabled us to compare classifications based on different generalizations and traditional network measures on equal footing using node-centric learning framework.

Our results show that the decision tree classifier generally outperforms the logistic regression classifier. Our evaluation of each degree generalization individually show that neither the *C-degree* nor the *α-degree* has consistence performance over the other generalization or the unweighted degree measure over all datasets using different classifiers. However, the *C-degree* generalization outperforms the *α-degree* in overall in more datasets than the *α-degree* does, if we choose the best classifier. This finding may be due to the reliance of *α-degree* generalization on the tuning parameter α without clear stated setup guidelines or effect of edge weights, unlike the *C-degree* methodology which is parameterless and depends on the value of each edge weight. Notably, the joint use of nodes *C-degree* with strength records consistent improvement in efficiency over the *α-degrees*, the traditional unweighted degree and the strength measures. However, whether the joint use of the *C-degree* and the node's strength is considered as an important combination of measures and to what extent it will effectively improve classifications and complex networks mining, remains an open research question.

As a future work, we are interested in extending our comparative methodology to evaluate other generalization methodologies (such as the ensemble approach [2]) as well as other unweighted network measures (such as the clustering cooffecient). This suggests more studies to be performed to compare different generalizations, but with very careful selection of datasets or by examining these measures on small world networks to have more clear results. We have a strong belief based on solid theoretical proofs and calculations that the generalizations based on effective cardinality such as the *C-degree* can dominate other generalized measures if applied to networks with weights disparity. Therefore, further comparisons and evaluation studies will be of great importance in complex networks mining researches.

Appendix A

Implementation

Before starting the implementation of our methodology, we spent a lot of time to get familiar with the NetKit-SRL[1] software (modules structure) in order to know where to implement the different network measures functions, and how to construct classifiers and generate classification results. The modularity of the toolkit and the plug and play architecture eases our job of introducing new java components and customizing the configurations.

A.1 Aggregators

To accommodate our goal of evaluating different generalizations, we implement five aggregators to compute the addressed network measures: *C-degree*, α 1.5 degree, α 0.5 degree, node's degree and node's strength (as shown in table A.1), where the original NetKit offers seven properties aggregation components (aggregators): count, exist, max, mean, min, mode and ratio.

Aggregator	**Feature Vector**
NodeCdegree(Attribute, Value)	Node's C-degree
NodeAlphaDegree(Attribute, Value)	Node's α 0.5 Degree
NodeAlpha_Degree(Attribute, Value)	Node's α 1.5 Degree
NodeDegree(Attribute, Value)	Node's Degree
NodeStrength(Attribute, Value)	Node's Strength

Table A.1: Newly Implemented Network Measures Aggregators

We add the newly implemented java aggregators under aggregators package
$\backslash src \backslash netkit \backslash classifiers \backslash aggregators$

We use each of the newly implemented aggregators to construct two new Network-Only Link-Based classifiers, one with Logistic regression classifier and the other with J84 classifier (as shown in tables A.2 and A.3).

Classifier Name	**Feature Vector**
nolb-lr-nodecdegree	Node's C-degree
nolb-lr-nodealpha0.5	Node's α 0.5 degree
nolb-lr-nodealpha1.5	Node's α 1.5 degree
nolb-lr-nodedegree	Node's Degree
nolb-lr-nodestrength	Node's Strength

Table A.2: Constructed Network-only Link-Based relational classifier with Logistic

To study the effect of combining more than one network measure for data classification, we can override the configuration of any of the constructed classifiers to accept more than one aggregators using the command-line option (*-aggregators aggregator1,aggregator2*). For example, to create a vector

[1]NetKit-SRL User guide is available at http://www-bcf.usc.edu/ macskass/NetKit.html

Classifier Name	Feature Vector
nolb-j48-nodecdegree	Node's C-degree
nolb-j48-nodealpha0.5	Node's α 0.5 degree
nolb-j48-nodealpha1.5	Node's α 1.5 degree
nolb-j48-nodedegree	Node's Degree
nolb-j48-nodestrength	Node's Strength

Table A.3: Constructed Network-only Link-Based relational classifier with J48

of combined node's *C-degree* and strength measures we use the option *-aggregators Node-Cdegree,Node-Strength*.

In addition, we implement more five aggregators that are not reported in this thesis and can be used in future studies such as node's clustering coefficient, node's continuous clustering coefficient (see table A.4).Moreover we add other WEKA classifier such as *SMO* and *IBK* but results are not reported here in this thesis.

Aggregator	Feature Vector
ClassCdegree(Attribute, Value)	Class's C-degree
ClassClusteringCoeff(Attribute, Value)	Class's Clustering Coefficient
ClassCClusteringCoeff(Attribute, Value)	Class's Continious Clustering Coefficient
ClassDegree(Attribute, Value)	Class's Degree
ClassStrength(Attribute, Value)	Class's Strength

Table A.4: Other Implemented Network Measures Aggregators

A.2 Aggregators and Classifiers Configurations

To get new components working, we should configure them in terms of name, class and attributes setup, where Netkit-SRL has nine configuration files. To customize Netkit-SRL we edit the configuration files for both aggregators and relational classifiers.

We define the new aggregators of network measures in aggregators properties file ($\backslash lib\backslash aggregator.properties$) as shown in table A.5.

% Node's C-degree: NodeCdegree(Attribute, Value) NodeCdegree.class=netkit.classifiers.aggregators.NodeCdegree NodeCdegree.accept=CATEGORICAL,DISCRETE
% Node's α 0.5 Degree: NodeAlphaDegree(Attribute, Value) NodeAlphaDegree.class=netkit.classifiers.aggregators.NodeAlphaDegree NodeAlphaDegree.accept=CATEGORICAL,DISCRETE
% Node's α 1.5 Degree: NodeAlpha_Degree(Attribute, Value) NodeAlpha_Degree.class=netkit.classifiers.aggregators.NodeAlpha_Degree NodeAlpha_Degree.accept=CATEGORICAL,DISCRETE
% Node's Degree: NodeDegree(Attribute, Value) NodeDegree.class=netkit.classifiers.aggregators.NodeDegree NodeDegree.accept=CATEGORICAL,DISCRETE
% Node's Strength: Node-Strength(Attribute, Value) NodeStrength.class=netkit.classifiers.aggregators.NodeStrength NodeStrength.accept=CATEGORICAL,DISCRETE

Table A.5: Newly Implemented Aggregators Configuration

We customize, name and define, the constructed relational classifier in relational classifier properties file ($\backslash lib\backslash rclassifier.properties$) (as sample seen in table A.6).

% nLB Classifier: Node's C-degree with Logistic regression nolb-lr-nodecdegree.class=netkit.classifiers.relational.NetworkWeka nolb-lr-nodecdegree.classifier=logistic nolb-lr-nodecdegree.useintrinsic=false nolb-lr- nodecdegree.aggregators=Node-Cdegree nolb-lr-nodecdegree.aggregation=ClassOnly
% nLB Classifier: Node's C-degree with J48 nolb-j48-nodecdegree.class=netkit.classifiers.relational.NetworkWeka nolb-j48-nodecdegree.classifier=j48 nolb-j48-nodecdegree.useintrinsic=false nolb-j48-nodecdegree.aggregators=Node-Cdegree nolb-j48-nodecdegree.aggregation=ClassOnly

Table A.6: Sample of Network-only Link-Base Classifier Configurations

Appendix B

Classification Experiments Runs Using Logistic Regression

RUN	D	CD	S	CD&S	D&S	α 0.5 Degree	α 1.5 Degree
1	0.5455	0.4242	0.4848	0.6364	0.5152	0.5758	0.5152
2	0.5588	0.5294	0.5294	0.5294	0.5882	0.5588	0.5588
3	0.5588	0.5000	0.5294	0.6176	0.5882	0.6471	0.5588
4	0.6471	0.6471	0.5588	0.7059	0.5882	0.6176	0.4412
5	0.6471	0.5882	0.5000	0.5588	0.5882	0.6176	0.5294
6	0.5758	0.5588	0.6061	0.6970	0.54545	0.6061	0.4848
7	0.6176	0.5152	0.6471	0.6176	0.5294	0.6176	0.6176
8	0.6471	0.5294	0.5882	0.6471	0.5882	0.5294	0.6176
9	0.6176	0.5882	0.5882	0.6176	0.5294	0.5882	0.5000
10	0.6176	0.6765	0.6471	0.5588	0.6177	0.6176	0.5882
Final Accuracy	0.6033	0.5557	0.5679	0.6186	0.5976	0.5412	0.5678

Table B.1: Logistic Regression Classifier: WebKB-Texas-cocite Classification Experiments Runs

RUN	D	CD	S	CD&S	D&S	α 0.5 Degree	α 1.5 Degree
1	0.6744	0.5814	0.6744	0.6977	0.7209	0.6279	0.6279
2	0.6744	0.5116	0.6512	0.6977	0.6744	0.5581	0.6279
3	0.5455	0.6047	0.5909	0.5909	0.6364	0.6591	0.6136
4	0.6279	0.5227	0.6977	0.7209	0.6047	0.6977	0.6047
5	0.6591	0.5000	0.6364	0.6818	0.6591	0.6136	0.5227
6	0.5349	0.5581	0.6512	0.5814	0.6744	0.6047	0.6279
7	0.7209	0.5349	0.6512	0.6977	0.6279	0.6512	0.6512
8	0.5682	0.5000	0.6136	0.5000	0.6591	0.6136	0.5682
9	0.6977	0.6047	0.6977	0.6047	0.6512	0.6744	0.5814
10	0.6977	0.5814	0.5814	0.5814	0.5349	0.6512	0.5814
Final Accuracy	0.6400	0.5499	0.6445	0.6354	0.6351	0.6007	0.6443

Table B.2: Logistic Regression Classifier: WebKB-Washington-cocite Classification Experiments Runs

RUN	D	CD	S	CD&S	D&S	α 0.5 Degree	α 1.5 Degree
1	0.4884	0.4884	0.5116	0.5581	0.7429	0.4419	0.4186
2	0.4884	0.4186	0.3488	0.5349	0.6286	0.4419	0.5814
3	0.5000	0.5682	0.5000	0.5682	0.6388	0.4545	0.4318
4	0.4884	0.5581	0.3953	0.4186	0.6571	0.4884	0.4651
5	0.5455	0.6136	0.5682	0.5000	0.5277	0.5682	0.5682
6	0.3953	0.6279	0.4186	0.5581	0.6857	0.4651	0.4884
7	0.5116	0.4651	0.3488	0.5814	0.6857	0.4884	0.3953
8	0.4545	0.5909	0.4773	0.5682	0.6111	0.4091	0.4091
9	0.4884	0.6512	0.6279	0.6279	0.6	0.4884	0.3721
10	0.6047	0.5116	0.5116	0.6512	0.5143	0.4419	0.4419
Final Accuracy	0.4965	0.5494	0.4708	0.5567	0.4688	0.4571	0.6292

Table B.3: Logistic Regression Classifier: WebKB-Washington-Link Classification Experiments Runs

RUN	D	CD	S	CD&S	D&S	α 0.5 Degree	α 1.5 Degree
1	0.7429	0.6286	0.7143	0.6000	0.7429	0.7143	0.6286
2	0.6286	0.5429	0.6000	0.6857	0.6286	0.6286	0.5714
3	0.6667	0.5556	0.6389	0.7778	0.6388	0.6944	0.6389
4	0.6857	0.5714	0.5429	0.4857	0.6571	0.6000	0.4857
5	0.6667	0.6286	0.6667	0.6944	0.5277	0.6111	0.6389
6	0.6286	0.6600	0.6857	0.7429	0.6857	0.6571	0.6000
7	0.6571	0.6000	0.6571	0.7143	0.6857	0.6571	0.5714
8	0.6667	0.5833	0.5833	0.5833	0.6111	0.6389	0.5833
9	0.6857	0.5714	0.6857	0.6286	0.6	0.6857	0.4857
10	0.6286	0.6286	0.5429	0.5429	0.5143	0.6286	0.5143
Final Accuracy	0.6657	0.5970	0.6317	0.6455	0.6516	0.5718	0.6292

Table B.4: Logistic Regression Classifier: WebKB-Wisconsin-cocite Classification Experiments Runs

RUN	D	CD	S	CD&S	D&S	α 0.5 Degree	α 1.5 Degree
1	0.6736	0.6806	0.7083	0.6806	0.7153	0.7847	0.6944
2	0.7431	0.6319	0.6944	0.6319	0.7153	0.7153	0.6667
3	0.7500	0.6944	0.6806	0.6944	0.7083	0.7083	0.7569
4	0.7222	0.6736	0.7014	0.6736	0.7431	0.7292	0.7014
5	0.7708	0.7431	0.6389	0.7431	0.6597	0.7500	0.5972
6	0.7222	0.7292	0.7431	0.7292	0.6944	0.6944	0.7222
7	0.7222	0.7361	0.6597	0.7361	0.6875	0.7639	0.6389
8	0.6597	0.7014	0.7569	0.7014	0.7153	0.6042	0.6875
9	0.6806	0.7778	0.6806	0.7778	0.6458	0.7292	0.6806
10	0.7431	0.7569	0.7222	0.7569	0.7222	0.6806	0.7083
Final Accuracy	0.7431	0.7569	0.7222	0.7569	0.6806	0.7083	0.7222

Table B.5: Logistic Regression Classifier: IMDb-all Classification Experiments Runs

RUN	D	CD	S	CD&S	D&S	α 0.5 Degree	α 1.5 Degree
1	0.3028	0.3853	0.3211	0.3716	0.3303	0.3119	0.3257
2	0.2740	0.3014	0.3105	0.3744	0.2922	0.2922	0.3014
3	0.2740	0.2922	0.2922	0.3653	0.2922	0.3059	0.2831
4	0.2648	0.3242	0.2968	0.3607	0.2785	0.2968	0.2877
5	0.2603	0.3333	0.2785	0.3562	0.2740	0.2831	0.2831
6	0.2922	0.3151	0.3059	0.3836	0.3014	0.2968	0.3333
7	0.2648	0.3425	0.2968	0.3836	0.2922	0.2740	0.2922
8	0.2740	0.4018	0.2968	0.3744	0.2922	0.2922	0.2922
9	0.2740	0.3744	0.2831	0.3242	0.3105	0.2968	0.2968
10	0.2603	0.3151	0.2785	0.3196	0.2922	0.2694	0.2785
Final Accuracy	0.2741	0.3385	0.2960	0.3614	0.2919	0.2974	0.2956

Table B.6: Logistic Regression Classifier: Industry-pr Classification Experiments Runs

RUN	D	CD	S	CD&S	D&S	α 0.5 Degree	α 1.5 Degree
1	0.3128	0.3408	0.2793	0.3631	0.2793	0.2793	0.2793
2	0.2833	0.3111	0.2833	0.3167	0.2833	0.2833	0.2833
3	0.2778	0.3667	0.2778	0.3611	0.277	0.2778	0.2778
4	0.2833	0.3944	0.2833	0.3722	0.2833	0.2833	0.2833
5	0.2778	0.3778	0.2778	0.3333	0.277	0.2778	0.2778
6	0.2235	0.4190	0.2849	0.3799	0.2849	0.2849	0.2849
7	0.2778	0.3222	0.2778	0.3778	0.277	0.2778	0.2778
8	0.2833	0.3056	0.2833	0.3667	0.2833	0.2833	0.2833
9	0.2778	0.3611	0.2778	0.3389	0.277	0.2778	0.2778
10	0.2849	0.3520	0.2849	0.3631	0.2849	0.2849	0.2849
Final Accuracy	0.2782	0.3551	0.2810	0.3573	0.2810	0.2810	0.2810

Table B.7: Logistic Regression Classifier: Industry-yh Classification Experiments Runs

Appendix C

Classification Experiments Runs Using J48

RUN	D	CD	S	CD&S	D&S	α 0.5 Degree	α 1.5 Degree
1	0.8485	0.6667	0.6667	0.6970	0.7273	0.6970	0.5152
2	0.6471	0.7059	0.5882	0.7059	0.7647	0.7059	0.4706
3	0.6471	0.7941	0.6176	0.6765	0.8235	0.6765	0.5294
4	0.6765	0.6471	0.4706	0.7059	0.7647	0.6765	0.5882
5	0.6176	0.6765	0.6765	0.7059	0.6765	0.6765	0.6176
6	0.6970	0.7273	0.6061	0.6667	0.7879	0.7353	0.6364
7	0.6765	0.6765	0.6471	0.6765	0.7059	0.7273	0.5882
8	0.7941	0.7941	0.6471	0.8235	0.6176	0.6765	0.5882
9	0.7941	0.6765	0.5588	0.8235	0.8235	0.5882	0.5294
10	0.7059	0.7647	0.5882	0.7059	0.6765	0.5882	0.7059
Final Accuracy	0.7104	0.7129	0.6067	0.7187	0.6748	0.5769	0.7368

Table C.1: J48 Classifier: WebKB-Texas-cocite Classification Experiments Runs

RUN	D	CD	S	CD&S	D&S	α 0.5 Degree	α 1.5 Degree
1	0.8140	0.6977	0.6744	0.7442	0.6512	0.7674	0.6512
2	0.6279	0.5581	0.6279	0.6047	0.6512	0.6512	0.6279
3	0.6818	0.6136	0.5909	0.6136	0.6364	0.6364	0.5682
4	0.5814	0.5814	0.6047	0.6744	0.6047	0.6744	0.6512
5	0.5909	0.7045	0.5227	0.5000	0.7045	0.6136	0.6591
6	0.6512	0.5581	0.6279	0.6279	0.6279	0.6977	0.6512
7	0.6047	0.5581	0.6744	0.6744	0.7442	0.6279	0.6512
8	0.6818	0.7500	0.5909	0.5682	0.4773	0.5455	0.5909
9	0.6977	0.6047	0.6744	0.5116	0.6977	0.6977	0.5581
10	0.6512	0.7442	0.6512	0.6512	0.6744	0.6279	0.6279
Final Accuracy	0.6582	0.6371	0.6239	0.6170	0.6540	0.6238	0.6469

Table C.2: J48 Classifier: WebKB-Washington-cocite Classification Experiments Runs

RUN	D	CD	S	CD&S	D&S	α 0.5 Degree	α 1.5 Degree
1	0.4884	0.5814	0.4186	0.4651	0.5349	0.5349	0.3721
2	0.5116	0.7209	0.4884	0.6512	0.5814	0.4884	0.4651
3	0.4318	0.6591	0.5000	0.4773	0.5455	0.5000	0.4091
4	0.5581	0.5814	0.5349	0.6512	0.4884	0.6047	0.5349
5	0.5000	0.6136	0.3182	0.6364	0.5455	0.3636	0.4545
6	0.4186	0.5814	0.6047	0.5581	0.3721	0.4419	0.5814
7	0.5581	0.5349	0.5581	0.6744	0.4186	0.5349	0.5349
8	0.4091	0.6364	0.4545	0.6364	0.5682	0.5227	0.4318
9	0.4884	0.6047	0.4884	0.6047	0.4651	0.4186	0.5814
10	0.5349	0.6279	0.6279	0.4884	0.4884	0.4884	0.5349
Final Accuracy	0.4899	0.6142	0.4994	0.5843	0.4898	0.4900	0.5008

Table C.3: J48 Classifier: WebKB-Washington-Link Classification Experiments Runs

RUN	D	CD	S	CD&S	D&S	α 0.5 Degree	α 1.5 Degree
1	0.7429	0.6286	0.6571	0.7143	0.6857	0.7714	0.6571
2	0.7143	0.6571	0.6571	0.6857	0.6571	0.6857	0.6000
3	0.7222	0.6944	0.6389	0.6944	0.6667	0.6389	0.6389
4	0.6857	0.5143	0.6571	0.6000	0.6571	0.6571	0.6000
5	0.6389	0.6944	0.6667	0.7222	0.7778	0.7500	0.6667
6	0.6286	0.6286	0.6857	0.6857	0.7143	0.6286	0.6571
7	0.7429	0.7429	0.6000	0.6857	0.7429	0.7143	0.6571
8	0.7500	0.6667	0.6389	0.6389	0.6111	0.5278	0.6389
9	0.6857	0.6286	0.6286	0.6000	0.6857	0.6000	0.6571
10	0.6857	0.6571	0.5714	0.7143	0.6857	0.6571	0.6857
Final Accuracy	0.6997	0.6513	0.6402	0.6741	0.6631	0.6459	0.6884

Table C.4: J48 Classifier: WebKB-Wisconsin-cocite Classification Experiments Runs

RUN	D	CD	S	CD&S	D&S	α 0.5 Degree	α 1.5 Degree
1	0.7639	0.6944	0.6528	0.6319	0.7222	0.7222	0.7569
2	0.6458	0.6667	0.6806	0.6944	0.6944	0.7222	0.7639
3	0.6875	0.7847	0.6736	0.7292	0.7083	0.6597	0.7639
4	0.6944	0.7014	0.7014	0.7222	0.6944	0.7292	0.6875
5	0.7292	0.6944	0.7222	0.7083	0.7153	0.7917	0.7431
6	0.7153	0.6181	0.7014	0.7361	0.6597	0.7500	0.6458
7	0.6528	0.7014	0.7569	0.6042	0.7569	0.6736	0.6736
8	0.7222	0.6944	0.7639	0.6528	0.7431	0.7083	0.7431
9	0.6458	0.6736	0.7431	0.6944	0.7431	0.6944	0.6597
10	0.7639	0.7153	0.6944	0.6806	0.6806	0.7083	0.6319
Final Accuracy	0.7021	0.6944	0.7090	0.6854	0.7160	0.7069	0.7118

Table C.5: J48 Classifier: IMDb-all Classification Experiments Runs

RUN	D	CD	S	CD&S	D&S	α 0.5 Degree	α 1.5 Degree
1	0.3211	0.4404	0.3165	0.4725	0.2890	0.2661	0.3119
2	0.2831	0.3653	0.3105	0.4292	0.2648	0.2740	0.3333
3	0.3059	0.4155	0.2968	0.3379	0.2785	0.2603	0.2922
4	0.2922	0.3242	0.2922	0.4338	0.3014	0.2648	0.2466
5	0.3059	0.4110	0.2603	0.4110	0.2694	0.2420	0.2603
6	0.3151	0.4018	0.2831	0.4110	0.2922	0.3333	0.2922
7	0.2740	0.3973	0.2922	0.4110	0.2740	0.2785	0.3014
8	0.2648	0.3836	0.2877	0.4384	0.2648	0.2831	0.2648
9	0.2785	0.4155	0.2831	0.4018	0.3516	0.2831	0.2740
10	0.2785	0.3881	0.2831	0.3516	0.2694	0.2694	0.2877
Final Accuracy	0.2919	0.3943	0.2906	0.4098	0.2755	0.2864	0.2855

Table C.6: J48 Classifier: Industry-pr Classification Experiments Runs

RUN	D	CD	S	CD&S	D&S	α 0.5 Degree	α 1.5 Degree
1	0.2570	0.3073	0.2793	0.2961	0.2235	0.2905	0.2346
2	0.2611	0.2944	0.2222	0.2444	0.2611	0.2056	0.2000
3	0.2944	0.3111	0.2333	0.2889	0.2500	0.2389	0.2278
4	0.2444	0.2444	0.2333	0.3000	0.2778	0.2167	0.2500
5	0.3056	0.2556	0.2333	0.3722	0.2500	0.2778	0.2889
6	0.2458	0.3520	0.2235	0.3520	0.2905	0.2458	0.2458
7	0.3111	0.3278	0.2667	0.3111	0.2778	0.2556	0.2333
8	0.2333	0.2889	0.2722	0.3444	0.2667	0.2333	0.2222
9	0.2667	0.2944	0.2333	0.3111	0.2389	0.2333	0.2556
10	0.3073	0.3017	0.2682	0.3464	0.2179	0.2737	0.2458
Final Accuracy	0.2727	0.2978	0.2465	0.3167	0.2471	0.2404	0.2554

Table C.7: J48 Classifier: Industry-yh Classification Experiments Runs

Bibliography

[1] Abdallah, S. 2011, Generalizing unweighted network measures to capture the focus in interactions, Springer Wien, pp. 1–15. 10.1007/s13278-011-0018-8.
URL: *http://dx.doi.org/10.1007/s13278-011-0018-8*

[2] Ahnert, S. E., Garlaschelli, D., Fink, T. M. A. and Caldarelli, G. 2007, "Ensemble approach to the analysis of weighted networks", *Phys Rev E* , Vol. 76, APS.
URL: *http://dx.doi.org/10.1103/PhysRevE.76.016101*

[3] Barabasi, A. L. and Albert, R. 1999, "Emergence of scaling in random networks", *Science* , Vol. 286, Department of Physics, University of Notre Dame, Notre Dame, IN 46556, USA., pp. 509–512.
URL: *http://view.ncbi.nlm.nih.gov/pubmed/10521342*

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

[5] Barthélemy, M., Barrat, A., Pastor-Satorras, R. and Vespignani, A. 2005, "Characterization and modeling of weighted networks", *Physica A* , Vol. 346, pp. 34–43.

[6] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D.-U. 2006, "Complex networks: Structure and dynamics", *Phys. Rep.* , Vol. 424, pp. 175–308.

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

[8] Costa, L. d. F., Oliveira Jr., O. N., Travieso, G., Rodrigues, F. A., Villas Boas, P. R., Antiqueira, L., Viana, M. P. and Rocha, L. E. C. 2008, "Analyzing and modeling real-world phenomena with complex networks: A survey of applications".
URL: *http://arxiv.org/abs/0711.3199*

[9] Costa, L. d. F., Rodrigues, F. A., Travieso, G. and Villas Boas, P. R. 2006, "Characterization of complex networks: A survey of measurements", *Advances in Physics* , Vol. 56, Taylor & Francis, pp. 167–242.
URL: *http://dx.doi.org/10.1080/00018730601170527*

[10] Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K. and Slattery, S. 1998, Learning to extract symbolic knowledge from the world wide web, *in* 'Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence', AAAI '98/IAAI '98, American Association for Artificial Intelligence, Menlo Park, CA, USA, pp. 509–516.
URL: *http://portal.acm.org/citation.cfm?id=295240.295725*

[11] Desrosiers, C. and Karypis, G. 2009, Within-network classification using local structure similarity, *in* 'Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I', ECML PKDD '09, Springer-Verlag, Berlin, Heidelberg, pp. 260–275.
URL: *http://dx.doi.org/10.1007/978-3-642-04180-8_34*

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

[13] Heatherly, R., Kantarcioglu, M. and Thuraisingham, B. 2009, Social network classification incorporating link typevalues, *in* 'Proceedings of the 2009 IEEE international conference on Intelligence and security informatics', ISI'09, IEEE Press, Piscataway, NJ, USA, pp. 19–24.
URL: *http://portal.acm.org/citation.cfm?id=1706428.1706432*

[14] Jensen, D. and Neville, J. 2003, "Data mining in social networks", *Dynamic Social Network Modeling and Analysis Workshop Summary and Papers* , In National Academy of Sciences Symposium on Dynamic Social Network Modeling and Analysis, pp. 287–302.
URL: *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.63.6148*

[15] Kalisky, T., Sreenivasan, S., Braunstein, L., Buldyrev, S., Havlin, S. and H., S. 2006, "Scale-free networks emerging from weighted random graphs", *Phys. Rev. E* , Vol. 73, APS, p. 025103.
URL: *http://link.aps.org/abstract/PRE/v73/e025103*

[16] Lu, Q. and Getoor, L. 2003, Link-based classification, *in* T. Fawcett, N. Mishra, T. Fawcett and N. Mishra, eds, 'ICML', AAAI Press, pp. 496–503.

[17] Macskassy, S. and Provost, F. 2007, "Classification in networked data: A toolkit and a univariate case study", *J. Mach. Learn. Res.* , Vol. 8, MIT Press, Cambridge, MA, USA, pp. 935–983.

[18] Mcdowell, L. K., Gupta, K. M. and Aha, D. W. 2007, Case-based collective classification, *in* 'In Proceedings of the Twentieth International FLAIRS Conference. Key West, FL: AAAI'.

[19] Neville, J. and Jensen, D. 2003, "Collective classification with relational dependency networks", *Journal of Machine Learning Research* , Vol. 8, p. 2007.

[20] Newman, M. E. 2003, "The structure and function of complex networks", *SIAM Review* , Vol. 45, pp. 167–256.
URL: *http://arxiv.org/abs/cond-mat/0303516v1*

[21] Newman, M. E. 2004, "Analysis of weighted networks", *Phys. Rev. E* , Vol. 70, p. 056131.

[22] Newman, M. E. 2006, "Finding community structure in networks using the eigenvectors of matrices", *Phys. Rev. E* , Vol. 74, APS, p. 036104.

[23] Opsahl, T., Agneessens, F. and Skvoretz, J. 2010, "Node centrality in weighted networks: Generalizing degree and shortest paths", *Social Networks* , Vol. 32, pp. 251–245.

[24] Opsahl, T. and Panzarasa, P. 2009, "Clustering in weighted networks", *Social Networks* , Vol. 31, pp. 155–163.
URL: *http://dx.doi.org/10.1016/j.socnet.2009.02.002*

[25] Park, J. and Barabási, A.-L. n.d., "Distribution of node characteristics in complex networks.", *Proceedings of the National Academy of Sciences of the United States of America* , Vol. 104, pp. 17916–17920.
URL: *http://dx.doi.org/10.1073/pnas.0705081104*

[26] Taskar, B., Abbeel, P. and Koller, D. 2002, Discriminative probabilistic models of relational data, *in* 'Uncertainty in Artificial Intelligence', pp. 485–492.

[27] Taskar, B., Wong, M. F., Abbeel, P. and Koller, D. 2003, Link prediction in relational data, *in* 'in Neural Information Processing Systems'.
URL: *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.9.3546*

[28] Watts, D. J. and Strogatz, S. H. 1998, "Collective dynamics of 'small-world' networks", *Nature* , Vol. 393, pp. 440–442.

[29] Witten, I. H. and Frank, E. 2000, *Data mining: practical machine learning tools and techniques with Java implementations*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.