DATA PREDICTION AND RECALCULATION OF MISSING DATA IN SOFT SET

MUHAMMAD SADIQ KHAN

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

DATA PREDICTION AND RECALCULATION OF MISSING DATA IN SOFT SET

MUHAMMAD SADIQ KHAN

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

UNIVERSITY OF MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Muhammad Sadiq Khan

Matric No: WHA140010

Name of Degree: PhD

Title of ~~Project/Research Report/Dissertation/~~Thesis ("This Work"): Data Prediction and Recalculation of Missing Data in Soft Set

Field of Study: Information Security

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

DATA PREDICTION AND RECALCULATION OF MISSING DATA IN SOFT SET

ABSTRACT

Uncertain data cannot be processed by using the regular tools and techniques of clear data. Special techniques like fuzzy set, rough set, and soft set need to be utilized when dealing with uncertain data, and each special technique comes with its own advantages and snags. Soft set is considered as the most appropriate of these techniques. A soft set application represents uncertain data in tabular form where all values are represented by 0 or 1. Researchers use soft set representation in a number of applications involving decision making, parameter reduction, medical diagnosis, and conflict analysis. Soft set binary data may be missing due to communicational errors or viral attacks etc. Soft sets with incomplete data cannot be used in applications.

Few researchers have worked on data filling and recalculating incomplete soft sets, and the current research focuses on predicting missing values and decision values from non-missing data or aggregates. A soft set needs to be preprocessed in order to obtain aggregates while no preprocessing is needed when aggregates are not required. Therefore, this research discusses the existing techniques in terms of preprocessed and unprocessed soft sets.

The currently available approaches in the preprocessed category recalculate partial missing data from aggregates, yet are unable to use the set of aggregates for recalculating entire values. This research presents a mathematical technique capable of recalculating overall missing values from available aggregates.

Also investigated are the techniques belonging to the unprocessed category, among them being DFIS, a novel data filling approach for an incomplete soft set, which seems to be the most suitable technique in handling incomplete soft set data. The result shows that DFIS possesses a persisting accuracy problem in prediction. DFIS predicts missing values through association between parameters, yet makes no distinction between the different associations. Thus, it ignores the role of the strongest association, which in turn results in low accuracy. This research rectifies this particular DFIS issue by using a new prediction technique through strongest association (PSA). The experimental result validates the high accuracy of PSA over DFIS after implementing both techniques in MATLAB and testing for data filling using bench mark data sets.

Further, this research applies PSA to online social networks (OSN) and detects a new kind of network community for those nodes that are associated with each other. The new network community is named ‘virtual community’ and the inter-associated nodes are named ‘prime nodes’. Researchers have found that the unavailability of complete OSN nodes results in a low accuracy of ranking algorithms. Therefore, this research predicts new links in two OSNs (Facebook and Twitter) data sets through association between prime nodes using PSA. By completing OSNs through association between prime nodes using PSA, this study demonstrates that the performance of famous ranking algorithms (k-Core and PageRank) can be significantly improved.

Keywords: Soft Set, Missing Data, Data Recalculation, Data Prediction, Link Prediction

RAMALAN DATA DAN PENGIRAAN SEMULA DATA HILANG DALAM SET LEMBUT

ABSTRAK

Data tidak-pasti tidak boleh diproses dengan menggunakan peralatan dan teknik yang sama digunakan untuk data jelas. Teknik-teknik khas seperti set kabur, set kasar, dan set lembut perlu digunakan apabila berurusan dengan data tidak-pasti, dan setiap teknik khas mempunyai kelebihan dan kekurangannya sendiri. Set lembut dianggap sebagai teknik yang paling sesuai dikalangan teknik-teknik khas ini. Aplikasi sesuatu set lembut mewakilkan data tidak-pasti dalam bentuk jadual di mana semua nilai diwakili oleh 0 atau 1. Para penyelidik menggunakan perwakilan set lembut dalam beberapa aplikasi yang melibatkan pembuatan keputusan, pengurangan parameter, diagnosis perubatan, dan analisis konflik. Data perduaan set lembut berkemungkinan hilang disebabkan kesilapan komunikasi atau serangan virus dan lain-lain. Set lembut dengan data yang tidak lengkap tidak boleh digunakan dalam aplikasi.

Beberapa penyelidik telah mengusahakan pengisian dan penghitungan data set lembut yang tidak lengkap, dan penyelidikan semasa member tumpuan kepada meramalkan nilai yang hilang dan nilai keputusan daripada data atau agregat yang lengkap. Sesuatu set lembut perlu diproses terlebih dahulu untuk mendapatkan agregat sementara tiada pra-pemprosesan diperlukan apabila agregat tidak diperlukan. Oleh itu, kajian ini membincangkan teknik-teknik sedia ada dalam bentuk set lembut yang menjalani pra-proses dan yang tidak diproses.

Pendekatan sedia ada dalam kategori pra-proses mengira semula separa data yang hilang daripada agregat, namun ianya tidak dapat menggunakan set agregat untuk menghitung semula nilai keseluruhan. Kajian ini membentangkan teknik matematik yang mampu mengira semula keseluruhan nilai hilang dari agregat yang tersedia.

Juga dikaji adalah teknik-teknik yang dimiliki oleh kategori tidak diproses, di antaranya ialah DFIS, suatu pendekatan pengisian data yang baru untuk set lembut yang tidak lengkap, yang merupakan teknik yang paling sesuai untuk mengendalikan set lembut idak lengkap. Hasilnya menunjukkan bahawa DFIS mempunyai masalah ketepatan dalam ramalan yang berterusan. DFIS meramalkan nilai-nilai yang hilang melalui hubungan antara parameter, namun tidak membezakan antara penyatuan yang berbeza. Oleh itu, ia mengabaikan peranan penyatuan terkuat, yang seterusnya menghasilkan ketepatan yang rendah. Kajian ini membetulkan isu DFIS dengan menggunakan teknik ramalan baru melalui penyatuan terkuat (PSA). Hasil eksperimen mengesahkan ketepatan tinggi PSA berbanding DFIS selepas kedua teknik dilaksanakan dalam MATLAB dan diuji dari segi pengisian data menggunakan set data piawai.

Selanjutnya, kajian ini menggunakan PSA untuk rangkaian sosial dalam talian (OSN) dan satu jenis komuniti rangkaian baru dikesan untuk nod-nod yang berkaitan diantara satu sama lain. Komuniti rangkaian baru ini dinamakan ‘komuniti maya’ dan nod yang berkaitan ini dinamakan ‘nod perdana’. Para penyelidik mendapati bahawa ketiadaan nod OSN yang lengkap menghasilkan ketepatan yang rendah untuk algoritma pemeringkatan. Oleh itu, kajian ini meramalkan hubungan baru dalam dua set data OSN (Facebook dan Twitter) melalui penyatuan antara nod perdana menggunakan PSA. Dengan melengkapkan OSN melalui penyatuan antara nod utama menggunakan PSA, kajian ini menunjukkan bahawa prestasi algoritma pemeringkatan yang terkenal (k-Core dan PageRank) dapat ditingkatkan dengan ketara.

Kata kunci: Set Lembut, Data Hilang, Kiraan Semula Data, Ramalan Data, Ramalan Pautan

ACKNOWLEDGEMENTS

I am most thankful to Almighty Allah for blessing me with everything, like the opportunity, time, wisdom, strength and ability for achieving this chellenging task.

I would like to express my sincere gratitude to my supervisors Dr. Tutut Herawan and Dr. Ainuddin Wahid Abdul Wahab for their continuous guidance, inspiration, support and encouragement. Their friendly support helped me to complete my research.

My heartiest thanks to my parents, siblings and friends who extended their support, encouragement, prayers and facilitations, specially to my father Maulana Muhammad Qari for his innumerable efforts, prayers and sacrifices for our successes.

My deepest gratitudes are to my wife Shaheen for her love, patience, sincerity, motivation and many sacrifices in this journey. My love and thanks to my sweet daughters Maryam, Zainab, Kalsoom and Rabia for their patience, love and cute support.

This work is dedicated to the most lovable person, my late mother Mahir Zuban whose dream was our education but she departed in the initial stages of our studies. May Allah bless her soul (Ameen)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

AT	:	Attribute
BIS	:	Boolean-valued Information System
Card	:	Cardinality
CD	:	Consistency Degree
CN	:	Consistency
Diag	:	Diagonal of table
EUH		Empty, Universal and Hybrid diagonals
ID	:	Inconsistency Degree
IN	:	Inconsistency
IND	:	Indiscernibility
Inf(i)	:	Influence of node i
LR	:	Left to Right
LUCAP	:	Lung Cancer set with Probes
Mod	:	Modulus
OSN	:	Online Social Network
PP	:	Pre Processed
PSA	:	Prediction through Strongest Association
RL	:	Right to Left
SPECT	:	Single Proton Emission Computed Tomography
Supp(u)	:	Supported values set for object u
U	:	Universal set
UP	:	Un Processed
$\forall$	:	For all
$\subseteq$	:	Is the subset of

E	:	imprecision function
c_i	:	Choice of object i
*	:	Unknown value
d_i	:	Decision value for object i
P_{bit}	:	Parity bit for row
C_{bit}	:	Parity bit for column
C_{agg}	:	Column aggregate
$\|U\|$	:	Absolute value of U
M_x	:	Spreading efficiency of x
Λ	:	Threshold lambda
$\Leftrightarrow$	:	Existence of association
$\nLeftrightarrow$	:	Existence of no association
$\Rrightarrow$	:	Inconsistent association

CHAPTER 1: INTRODUCTION

In this chapter, the rudimentary concepts of data types, clear data, uncertain and vague data, tools and techniques for handling vague data are briefly presented. Soft set theory, tabular representation of soft set and incomplete soft set are discussed in details.

1.1 Background

Facts and figures in pieces is called data or raw data, or information in such form that an entity (persons or organizations) cannot decide on its base without processing it further, or unprocessed information. After certain processing, raw data is converted into information. Processing of raw data depends on the requirement of processing entity, all entities process raw data in their own ways according to their own necessities for obtaining their desired outputs and decisions (Bellinger, Castro, & Mills, 2004).

A raw data X for an entity A can be information for another entity B at the same time. Because entity A needs it's further processing for obtaining their required output, while the same data can fulfill the requirement of entity B as the processed limit is sufficient for their needs. For example, the number of students in the language class is enough data for their language teacher but their attendance in all subjects including language class (further processed) is required for the examination section. After the entity B processes raw data from X to Y form and it becomes information for entity B, again this new data Y can be raw data for another entity C and so forth. In these cases, it can be seen that data X and Y are both information and raw data at the same time for different entities. Therefore, processed and unprocessed data (raw data and information) can be interchangeably used.

There are two main types of data called qualitative data and quantitative data. Qualitative data is obtained for getting knowledge, properties and qualities of things without involvement of numerical digits. Qualitative data is further divided to two sub-categories called nominal and ordinal. Nominal qualitative data is the one in which no pre-defined or standard structure exists rather everyone deal it according to his/her own requirements. Example of nominal qualitative data is the colors. White color of something can be white, light-white, full-white, cream-white, smoke-white and snow-white and so on. For ordinal qualitative data, a sequence is already defined in nature, it is used a as standard and no one can change it easily. For example, humans are generally categorized into male and female in term of gender. Quantitative data usually consists of numeric values and further divided into two sub-types known as discrete or integral and continuous or ratio quantitative data. Example of discrete data is number of students in language class; it must be in whole numbers, while continuous quantitative data can be described as the height of each of these students. Qualitative data can be converted or represented in quantitative forms as well, like, five black colors are represented by integers 1 to 5 as; dark-black = 1, light-black =2, bluish-black =3 reddish-black =4 and greenish-black =5. Some fuzziness or ambiguity or uncertainty in nature of data can be observed while looking at the example of different types of colors. Therefore, data is further divided into two other categories like crisp data and vague data.

1.2 Crisp data vs. unclear data

Crisp and unclear data is further explained below with examples.

1.2.1 Crisp data

Crisp data is also known as clear data or unambiguous data. The data which is clear, clean, and certain and has no ambiguity is called crisp data. For example; a university student’s database consists of student personal information like name, father name, addresses, nationality, contact info and previous education and university particulars like registration number, year of registration, current semester, previous performance, fee details, courses completed and current courses. In this example, data is certain, crisp and clear which contains no ambiguity and approximation in its processing. Although if processed through much complicated procedures, the answer and process is crisp and agreed among all, until the procedures used are valid and free from errors and mistakes. Such data have no ambiguity in processing (calculating) each student due fees, achieved percent marks etc. There are hundreds of kinds of crisp data in our daily life with hundred kinds of solutions in the form of mathematical theories, computer applications and research models.

1.2.2 Unclear data

In contrast to certain, unambiguous or crisp data, a lot of daily life problems in education, engineering, economics, social sciences, medical and computer science (artificial intelligence and cognitive sciences, especially in the areas of machine learning, knowledge acquisition, decision analysis, knowledge discovery from databases (KDD), expert systems, inductive reasoning and pattern recognition) encounter with data that have no crisp solution and no crisp representation if processed though ordinary crisp data tools and techniques (Kahraman, Onar, & Oztaysi, 2015). For example, birds (Penguins, bat?), tall man, beautiful women, creditworthy customer, responsible person, trusty friend. Processing vague data using improper tools and techniques may yield in extra-large, very small, unexpected and misleading results. Like crisp data, unclear data has also hundreds of kinds and its hundreds of proposed solutions for processing. Active research started in computer science, numerical analysis and mathematics on unclear data in early 1960s (Moore & Lodwick, 2003).

1.3 Tools and techniques used for handling unclear data

Prominent tools and techniques used for handling fuzzy data are based on the theories of probability, fuzzy set theory (L.A. Zadeh, 1965), rough set theory (Z. Pawlak, 1982), Intuitionistic fuzzy sets (Atanassov, 1986; Radicchi, Castellano, Cecconi, Loreto, & Parisi, 2004), Vague sets (Gau & Buehrer, 1993), theory of interval mathematics (Radicchi et al., 2004) and soft set theory (Molodtsov, 1999). Among them fuzzy set, rough set and soft set theories are most famed and they are overviewed below, one by one.

1.3.1 Fuzzy set theory

Let X is a universal set (objects/space of points) with its members x, i.e. $X = \{x\}$. A fuzzy set A in X is represented by characteristic function $f(x)$ such that $f(x)$ associates with each point of X through interval $[0, 1]$, X takes a real value in this interval for each of its membership association level e.g. $f(x) = 1$ if $x \in A$ and $f(x) = 0$ if $x \notin A$. Closer the value of x to 1 means higher grade of membership and closer the values of x to 0 means lower grade of membership e.g. we can have membership functions $f(x)$ of A as $f(1) = 0.03$, $f(2) = 0.21$, $f(3) = 0.17$, $f(101) = 0.77$, $f(996) = 0.84$ and $f(1000) = 1$(Lotfi A Zadeh, 1965; Zimmerman, 1991; H.-J. Zimmermann, 2001, 2014; H. Zimmermann, 1991).

In contrast to fuzzy set, the Ordinary set, crisp set or "set" takes only two values i.e. either 1 or 0 for completely belonging or completely not-belonging to X.

1.3.2 Rough set theory

According to this theory, each set of data can be represented in a set X of objects U having boundary lines called the lower approximation and upper approximation. The lower approximation and upper approximation are associated in a pair of crisp set such that the lower approximation consists of those objects which belongs to the set of data for sure while the upper approximation contains those objects which possibly belongs to the set of data and the difference between upper and lower approximation results in the boundary region of the data. The set X is called rough set if the boundary region has a non-empty value otherwise the set is crisp (non-vague) (Fortunato, 2010; Zdzisław Pawlak, 1982; Zdzislaw Pawlak, 1998; Z. Pawlak, 2012).

1.3.3 Soft set theory

Among previous theories of vague data, fuzzy set theory is most suitable because of its comparatively more mathematical presentation and natural look. But all have their own difficulties possibly due to their inadequacy in parameterization tools. Soft set theory is free from such difficulties because it uses adequate parameterization (Molodtsov, 1999).

Definition 1.1: Let U be a universal set and let E be a set of parameters then a pair (F,E) is called to be soft set over U if and only if F is a mapping of E into the set of all subsets of U

In other words, soft set is a parameterized family of the subsets of the set U. Every fuzzy set can be considered a special case of soft set.

1.3.3.1 Representation of soft set as a BIS (Standard Soft Set)

PK Maji used the concept of Yao and Lin (Lin, 1998; Yao, 1998) for representing soft set (F,E) in tabular form (P. Maji, Roy, & Biswas, 2002). According to this approach, all objects h_i of (F,E) are shown by rows and their parameters e_j by columns. For an object having certain parameter present i.e. $h_i \in F(e_j)$ is shown by putting its value equal to 1, otherwise zero as explained in below Example 1.1.

Example 1.1: Soft Set as BIS

Let $U = \{h_1, h_2, h_3, h_4, h_5, h_6\}$ be a set of houses and E = {expensive, beautiful, wooden, cheap, in the green surroundings, modern, in good repair, in bad repair} be a soft parameter. Consider the soft set (F, E) which describes the attractiveness of the houses, given by (F, E) = {Expensive houses $e_0 = \varphi$, beautiful houses $e_1 = \{h_1, h_2, h_3, h_4, h_5, h_6\}$, wooden houses $e_2 = \{h_1, h_2, h_6\}$, cheap houses $e_3 = \{h_1, h_2, h_3, h_4, h_5, h_6\}$, in the green surroundings houses $e_4 = \{h_1, h_2, h_3, h_4, h_6\}$, in good repair houses $e_5 = \{h_1, h_3, h_6\}$, modern houses $e_6 = \{h_1, h_2, h_6\}$, in bad repair houses $e_7 = \{h_2, h_4, h_5,\}$ }. (F, E) is represented in tabular form as shown in Table 1.1.

Table 1.1: Representation of Soft Set (F, E) in Tabular Form

$U\|E$	e_0	e_1	e_2	e_3	e_4	e_5	e_6	e_7
h_1	0	1	1	1	1	1	1	0
h_2	0	1	1	1	1	0	1	1
h_3	0	1	0	1	1	1	0	0
h_4	0	1	0	1	1	0	0	1
h_5	0	1	0	1	0	0	0	1
h_6	0	1	1	1	1	1	1	0

1.3.3.2 Applications of soft set theory

Soft set being represented in BIS Table 1.1 is applied in many applications. It is used for decision making and reduct in its initial application of representation in BIS (P. Maji et al., 2002). D Chen et al. redefined the reduct and showed that reduct and decision making presented by Maji is incorrect (Degang Chen, Tsang, Yeung, & Wang, 2005). Kong et al. showed that Chen et al. reduct can’t be applied to find sub-optimal choices and presented their technique of normal parameterization reduction technique which covers accuracy of sub-optimal choices as well (Kong, Gao, Wang, & Li, 2008). However, Kong et al. reduction technique is hard to understand and their reduction algorithm has high computational complexity. Ma et al. presented their technique of new efficient normal parameterization which is free from said difficulties (Qin, Ma, Herawan, & Zain, 2011a). Parameterization reduction in soft set is still an open problem and can be improved by presenting more efficient algorithms and new techniques.

Researchers extended soft set concept and applied it to different fields and daily life problems including medical diagnosis, data mining, and algebra.

1.3.3.3 Incomplete soft set:

Apart from hundreds of useful applications, sometimes the information or values of soft set gets missed due to security, data restriction, confidentiality, errors, mishandling, wrong entry or other possible reasons. In such cases, soft set with missing values becomes in incomplete. Incomplete soft set can no longer be used in lot of applications and if still used, might result in unexpected, wrong or very high or very less and misleading results.

Until now, few researchers have worked on handling with the situation of incomplete soft set. Initial work on incomplete soft set is data analysis approaches of soft sets under incomplete information (Zou & Xiao, 2008). This approach predicts only the decision or choice values in standard soft set using weighted average probability and the original missing values still remains missing. Data filling approach of soft set under incomplete information (DFIS) uses association between parameters to predict actual missing values in incomplete soft set and uses probability when there is no or weak association between parameters (Qin, Ma, Herawan & Zain, 2012a) A most recent approach, an efficient decision making approach in incomplete soft set improves the computational complexity of Zou et al approach and assign some values to originally missed values too (Kong et al., 2014). Other ways of handling incomplete soft set includes two techniques of re-calculating missing values from supported sets, parity bits and diagonals aggregates (Rose et al., 2011; Rose, Hassan, Awang, Herawan, & Deris, 2011).

1.4 Motivation

Data is the basic element for performing usual processing including most important operations of decision makings. Decision may be wrong if improper operations or tools are used for data processing, similarly the decision can be wrong if the data is not fully available, partially missing and/or improper technique is used for its prediction. Accurate data predictions have same importance as proper tools of data processing.

1.5 Problem statement

This research concluded from the literature, that existing techniques of handling incomplete soft set need to be categorized into two main types. First type of techniques relies on available values other than missing values (Kong et al., 2014; Qin et al., 2012a; Zou & Xiao, 2008). These techniques use association and probability to predict missing values. The results in this type of techniques are not 100% accurate and are improved gradually from one technique to another, either in term of accuracy, integrity and/or efficiency.

In contrast to first type, the second type of techniques (Mohd Rose et al., 2011; Rose et al., 2011) depend on the sets of equivalency in the form of aggregates as well as non-missing values. Missing data in this category is re-calculated from these equivalency sets and available values. The second type techniques don’t have the capability to re-calculate entire missing values from available aggregates.

Above stated limitations of both types of techniques indicates that accuracy improvement is an open problem in the first type of techniques and the techniques of second type can be extended to re-calculate overall missing values from available aggregates. Therefore, after categorization into two types, this research proposes an improved accuracy technique in one category and presents overall missing values re-calculation method from available aggregates in the other category.

1.6 Aim of the Research

The aim of this research is to study existing techniques of handling with incomplete soft sets, categorize them to two types and present new techniques that improve the accuracy and capability of both categories existing techniques.

1.7 Objectives

i. To investigate the accuracy and capability of techniques used for handling incomplete soft set and classify them in preprocessed and unprocessed categories

ii. To present a new concept in the preprocessed incomplete soft set category that is capable of re-calculating overall missing values from available aggregates

iii. To indicate the most suitable method in the unprocessed category of incomplete soft sets, find its weakness and improve its accuracy by presenting an alternative method

iv. To apply prediction of incomplete soft set though association to link prediction problem in Online Social Networks (OSNs)

1.8 Research Questions

To obtain objective of this research, the following questions need to answered

i. What is soft set, what are its applications, what is incomplete soft set and what are the techniques of handling missing data in soft set?

ii. How can the existing techniques in incomplete soft set be classified?

iii. Can the techniques of incomplete soft be used for re-calculating overall missing data from aggregates?

iv. Which existing data dependent technique is most suitable for predicting incomplete soft set values?

v. What is/are the drawback(s) of most suitable data dependent existing techniques and how they can be addressed?

vi. Can the association between parameter be applied to daily life problems like link prediction in OSNs?

1.9 Mapping of the Objectives with Research Questions

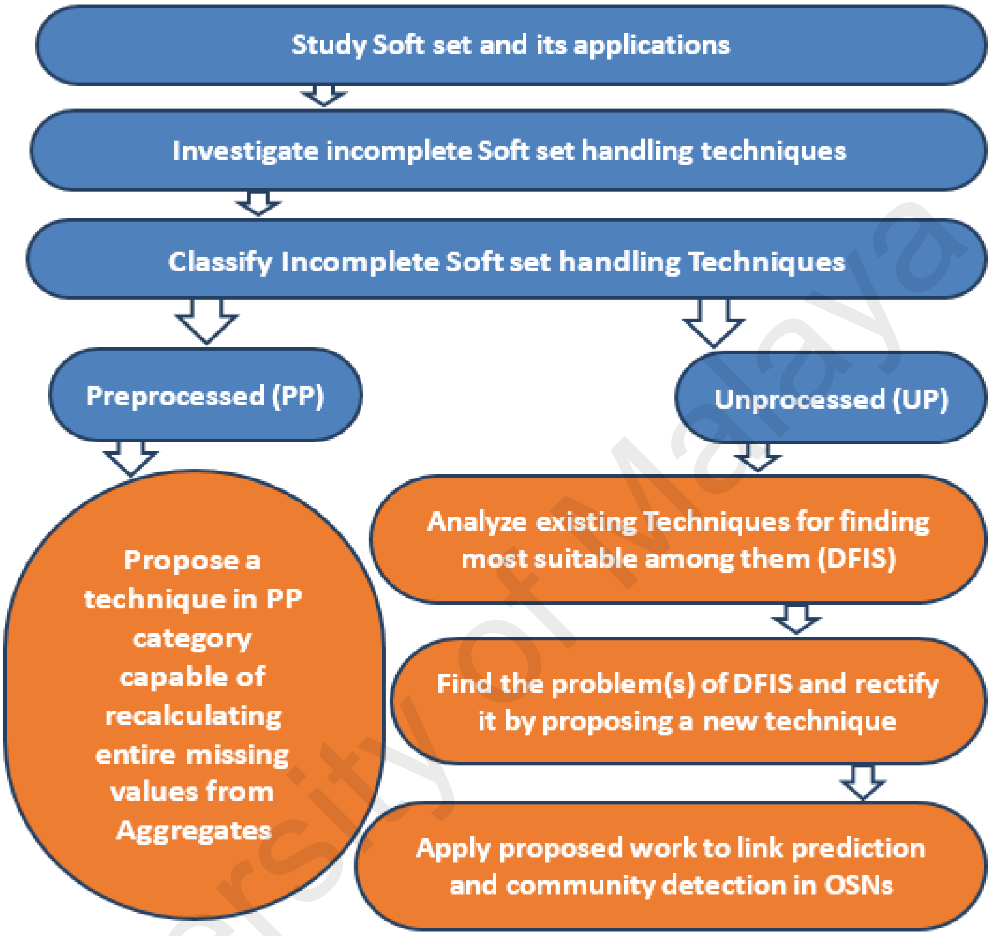
The mapping between objectives and research questions is provided in Table 1.2 to show how the research questions are connected with the objectives.

Table 1.2: Mapping of Objectives and Research Questions

Objectives	Research Questions
1. To investigate the accuracy and capability of techniques used for handling incomplete soft set and classify them in preprocessed and unprocessed categories	1. What is soft set, what are its applications, what is incomplete soft set and what are the techniques of handling missing data in soft set? 2. How can the existing techniques in incomplete soft set be classified?
2. To present a new concept in the preprocessed incomplete soft set category that is capable of re-calculating overall missing values from available aggregates	3. Can the techniques of incomplete soft be used for re-calculating overall missing data from aggregates?
3. To indicate the most suitable method in the unprocessed category of incomplete soft sets, find its weakness and improve its accuracy by presenting an alternative method	4. Which existing data depended technique is most suitable for predicting incomplete soft set values? 5. What is/are the drawback(s) of most suitable data dependent existing techniques and how they can be addressed?
4. To apply prediction of incomplete soft set though association to link prediction problem in Online Social Networks (OSNs)	6. Can the association between parameter be applied to daily life problems like link prediction in OSNs?

1.10 Methodology

In this section, the step by step procedures adopted to achieve the goals of this research are discussed. Methodology is summarized in a flow chart in Figure 1.1.



Study Soft set and its applications
Investigate incomplete Soft set handling techniques
Classify Incomplete Soft set handling Techniques
Preprocessed (PP)
Unprocessed (UP)
Propose a technique in PP category capable of recalculating entire missing values from Aggregates
Analyze existing Techniques for finding most suitable among them (DFIS)
Find the problem(s) of DFIS and rectify it by proposing a new technique
Apply proposed work to link prediction and community detection in OSNs

Figure 1.1: Methodology flow chart of the proposed study

Basic applications of soft set presented for parameterization reduction and decision making and the techniques used for handling incomplete soft in decision making are studied. The later techniques are further studied and categorized into two types based on data dependency and equivalency sets dependency parameters. It is shown that the techniques of one type depend on available data only while the other type techniques depend on equivalency sets as well.

First type of techniques can’t be used for recalculating overall missing values at all while the other type techniques also can’t be used in its current form to recalculate entire missing values from aggregates or equivalency sets. After this categorization, the techniques depending on equivalency sets are extended to be used for recalculating entire values from equivalency sets.

On the other hand, the techniques of other category (dependent on available data only) are analyzed and the most suitable technique among them is found in term of high accuracy, less computational complexity and maintaining integrity of soft set. The most suitable technique in this category uses association between parameters to predict missing values yet this technique ignores the weight of strongest association among all parameters and deal with all association equally. Due to this drawback, the accuracy of this technique is low and it is improved by addressing the said problem. The technique of existing approach is revised so that the weight of strongest associations is not ignored and unknowns are predicted through strongest association first. The proposed method in this category compares its accuracy with baseline by implementing both techniques in MATLAB and testing them for 4 UCI[1] benchmark and LUCAP[2] data sets.

Moreover, association between parameters is applied to link prediction problem in online social networks (OSNs) and a new kind of network community named as virtual community is identified through association between prime nodes. The new method of link prediction and virtual community detection is also implemented in MATLAB and new links are predicted through it for two real big data sets of global OSNs i.e. Facebook and Twitter. The results of proposed prediction are validated though well-

[1] UCI Machine Learning Repository 2013, https://archive.ics.uci.edu/ml/datasets.html. Accessed Dec 5, 2015

[2] Causality workbench 2013, http://www.causality.inf.ethz.ch/challenge.php?page=datasets. Accessed Dec 5, 2015

known ranking algorithms PageRank and k-Core by finding influential spreaders before and after links prediction.

1.11 Significance of the study

The first contribution of this thesis is recalculation of entire missing values from aggregates. This concept will open a new chapter for researchers in the development of novel applications in the fields of mathematics, especially in Boolean data, discrete mathematics, and computer science regardless of soft set or unclear data. It would be of great interest for mathematicians because it bypasses the restriction of solving simultaneous linear equation and has the capability to calculate more variables than available relations. This approach can be also applied to data novel compression at binary level in its future work.

The second contribution of this work is the data filling of partial missing values in soft set through strongest association between parameters. Soft set has been used in valuable applications like decision making and wrong or no decision can be made using missing data. Similarly, low accuracy of data used in decision making can result in wrong decision and wrong decisions can result in huge loss to organizations and individuals. As proposed approach has highest accuracy among all existing techniques therefore, most accurate decision making is expected using this technique for data filling.

The last contribution of this study is the application of proposed data prediction method in link prediction and new kind of community detection in OSNs. This work has direct significance to OSNs owners for their network growth. They can suggest new links of common interest to the “virtual community” members in their network recommender system and both users and network operating authorities can benefit from it.

1.12 Research contribution

Apart from classification of soft set handling techniques to PP and UP categories, this research has mainly two contributions i.e. recalculation of entire missing values from aggregates and data prediction through strongest association. Another third contribution comes from applying the data prediction through strongest association in link prediction problem in online social networks.

1.13 Organization of the thesis

The remaining of this thesis is organized as given below. This work contains 6 Chapters. Chapter wise description is discussed below and summarized in Figure 1.2.

1.13.1 Chapter 2

Basic applications of soft set are discussed in this chapter. A brief overview of general applications is discussed without going into details. More related works of decision making and parameterization reduction are discussed in detail examples. The techniques of incomplete soft set are comprehensively reviewed with examples in detail for their classification and analysis later in the related chapters. One of the contribution and application of proposed work is the link prediction in OSN and its validation through ranking algorithms, therefore, related work to link prediction and ranking algorithm is also presented in the end of this chapter.

1.13.2 Chapter 3

This is the first chapter of this study contributions and it has mainly two sub-contributions. Existing techniques of incomplete soft are analyzed in this chapter for classification into two categories UP and PP, first. The second contribution is related to PP category and a concept of entire missing values recalculation from aggregates in incomplete soft set is presented in this chapter. The proposed work is explained with the help of new definitions, algorithm and a solved example as a proof of concept.

1.13.3 Chapter 4

This is the second chapter of this study contributions related to UP category of classification. Existing techniques of this category are analyzed for indicating most appropriate technique among them and DFIS is indicated as same. Further investigated is the problem of DFIS with the help of available data in the literature and experiments and own experiments on benchmark data sets. An alternative data filling technique in incomplete soft is presented which operates on strongest association unlike DFIS. Both techniques (proposed and DFIS) are intercompared by implementing in MATLAB in testing for bench mark data sets. High accuracy of proposed work is presented and discussed with its shortcoming.

1.13.4 Chapter 5

This chapter is an application of proposed work, proposed in chapter 4. It is related to a new kind of network community detection in OSN through association between prime nodes and link prediction through it. Mathematical relations, definitions, algorithm and examples are presented for describing proposed application. New links are predicted using proposed work in Facebook and Twitter data sets. Results of PageRank and k-Core are intercompared for both data sets before and after prediction of new links. Improved accuracy in the results of ranking algorithms due to new links prediction is presented with necessary discussions.

1.13.5 Chapter 6

This chapter contains the conclusion and future direction of this work by reappraising the objectives. Main contributions of this thesis are summarized and future directions are proposed in this chapter.

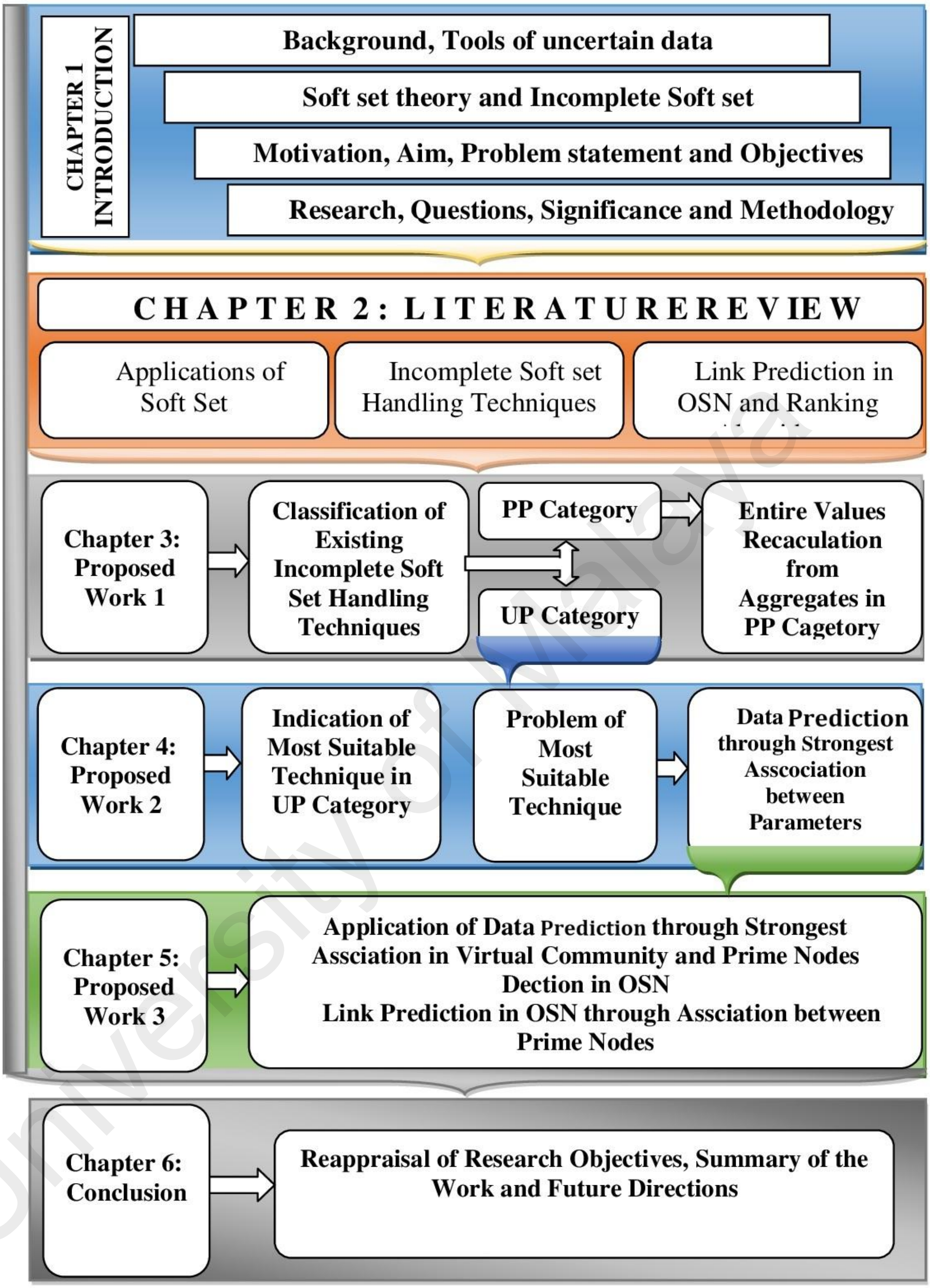

CHAPTER 1 INTRODUCTION
Background, Tools of uncertain data
Soft set theory and Incomplete Soft set
Motivation, Aim, Problem statement and Objectives
Research, Questions, Significance and Methodology
CHAPTER 2: LITERATUREREVIEW
Applications of Soft Set
Incomplete Soft set Handling Techniques
Link Prediction in OSN and Ranking
Chapter 3: Proposed Work 1
Classification of Existing Incomplete Soft Set Handling Techniques
PP Category
UP Category
Entire Values Recaculation from Aggregates in PP Cagetory
Chapter 4: Proposed Work 2
Indication of Most Suitable Technique in UP Category
Problem of Most Suitable Technique
Data Prediction through Strongest Asscociation between Parameters
Chapter 5: Proposed Work 3
Application of Data Prediction through Strongest Assciation in Virtual Community and Prime Nodes Dection in OSN
Link Prediction in OSN through Assciation between Prime Nodes
Chapter 6: Conclusion
Reappraisal of Research Objectives, Summary of the Work and Future Directions

Figure 1.2: Summary of thesis layout

CHAPTER 2: LITERRATURE REVIEW

This chapter is mainly divided into three parts, in first part: the major applications of soft set theory in decision making and parameter reduction are presented, the second part contains: the review of existing techniques for handling incomplete soft set in calculating decision values and predicting missing values, while link prediction and community detection techniques in online social networks and ranking algorithms are discussed in the third part. Link prediction in online social network and virtual community detection is an application of the UP category (UP category is discussed in chapter 4) of proposed work (proposed in chapter 5).

2.1 Applications of soft set theory

Since its presentation, the concept soft set theory has been applied in hundreds of commendable applications like medical diagnoses, decision making, artificial intelligence, soft computing, association rule mining, prediction, forecasting and many other fields. Few such applications of soft set are mentioned below.

Soft set theory (Ali, Feng, Liu, Min, & Shabir, 2009; P. Maji, Biswas, & Roy, 2003; Molodtsov, 1999) is applied in decision making and parameterization reduction (Çağman & Enginoğlu, 2010b; Degang Chen et al., 2005; Danjuma, Ismail, & Herawan, 2017; Isa, Rose, & Deris, 2011; Jiang, Liu, Tang, & Chen, 2011; Kong et al., 2008; P. Maji et al., 2002; P. K. Maji, 2012; Polat & Tanay, 2016; Qin et al., 2011a), in diagnoses of prostate cancer risk (Yuksel, Dizman, Yildizdan, & Sert, 2013), in association rules mining (Herawan & Deris, 2011), in decision making for patients suspected influenza-like illness (Herawan, 2010), in conflict analysis (Sutoyo, Mungad, Hamid, & Herawan, 2016).

Soft set is combined with other mathematical models. It is used in ideal theory of BCK/BCI-algebras and to ideals in d-algebras (Jun, Lee, & Park, 2009; Jun & Park, 2008). Lattice ordered soft sets are defined where the elements of parameters have some order (Ali, Mahmood, Rehman, & Aslam, 2015). Soft mapping is defined and applied to medical diagnosis (Majumdar & Samanta, 2010b). Soft-matrix is introduced and soft max-min decision making procedure is defined (Çağman & Enginoğlu, 2010a). Soft groups (Aktaş & Çağman, 2007), normalistic soft groups (Sezgin & Atagün, 2011), soft semirings (Feng, Jun, & Zhao, 2008) and algebraic structures of soft sets (Muhammad Irfan Ali, Shabir, & Naz, 2011) are defined. Soft set is extended to Soft β-Open Sets and Soft β-Continuous Functions (Akdag & Ozkan, 2014), Interval-valued vague soft sets (Alhazaymeh & Hassan, 2012), Soft expert sets (Alkhazaleh & Salleh, 2012), Multi aspect soft sets (Sulaiman & Mohamad, 2013), Neutrosophic soft set (P. K. Maji, 2013) and interval soft sets (X. Zhang, 2014).

To associate soft set with fuzzy set, the concept of fuzzy soft set and generalized fuzzy soft set (N Cagman, S Enginoglu, & F Citak, 2011; P. K. Maji, BISWAS, & Roy, 2001; Majumdar & Samanta, 2010a; X. Yang, Yu, Yang, & Wu, 2007) and intuitionistic fuzzy soft sets are introduced (P. K. Maji, 2009) and further contributions are made to fuzzy soft sets (Ahmad & Kharal, 2009). Fuzzy soft set is used in decision making (Alcantud, 2015, 2016; Alkhazaleh, 2015; Aslam & Abdullah, 2013; Basu, Mahapatra, & Mondal, 2012; Dinda, Bera, & Samanta, 2010; Feng, Jun, Liu, & Li, 2010; Kong, Gao, & Wang, 2009; Kong, Wang, & Wu, 2011; Z. Li, Wen, & Xie, 2015; Roy & Maji, 2007; Y. Yang, Tan, & Meng, 2013), its logic connectives are studied (Muhammad Irfan Ali & Shabir, 2014). Soft topological structure (Çağman, Karataş, & Enginoglu, 2011; Tanay & Kandemir, 2011), topological spaces are introduced (Aygünoğlu & Aygün, 2012; B. Chen, 2013; Hussain & Ahmad, 2011; Kannan, 2012; W. K. Min, 2011; Nazmul & Samanta, 2012; Shabir & Naz, 2011; Zorlutuna, Akdag, Min, & Atmaca, 2012) and combined recently with fuzzy set (Mahanta & Das, 2017). Intuitionistic fuzzy soft sets are used in decision making (Agarwal, Biswas, & Hanmandlu, 2013; Das & Kar, 2014; Deli & Karataş, 2016; Jiang, Tang, & Chen, 2011; Tripathy, Mohanty, & Sooraj, 2016; Z. Zhang, 2012). Interval-valued fuzzy soft sets (Jiang, Tang, Chen, Liu, & Tang, 2010) are defined and used in decision making (Feng, Li, & Leoreanu-Fotea, 2010).

Fuzzy soft lattices are defined and their structure is discussed (Shao & Qin, 2012). Hesitant fuzzy soft set is introduced and applied to decision making (Wang, Li, & Chen, 2014). Fuzzy soft set is also applied to diagnoses in medical (Çelik & Yamak, 2013) using fuzzy anathematic operations, to investment decision making problem (Kalaichelvi & Malini, 2011a), to forecasting approach (Xiao, Gong, & Zou, 2009), to flood prediction alarm (Kalayathankal & Suresh Singh, 2010). Researchers have also shown the association of soft set with rough set (Feng, 2009; Feng, Li, Davvaz, & Ali, 2010; Feng, Liu, Leoreanu-Fotea, & Jun, 2011; Herawan & Deris, 2009a; D. Pei & Miao, 2005) and vague soft set is extended from soft set (Xu, Ma, Wang, & Hao, 2010)

However, it is intolerable to discuss each of these applications in this work in details; therefore, most related applications of decision making and parameterization reduction are reviewed below.

Parameters reduction in soft set was initiated by PK Maji in his preliminary work (P. Maji et al., 2002), but there were some technical gaffes in his proposed algorithm of reduction which were gradually covered by Chen, Kong and Ma et al. in (Degang Chen et al., 2005; Kong et al., 2008; Qin et al., 2011a) respectively.

2.1.1 Application in deriving reduct table and decision making by PK Maji

PK Maji's reduction is based on his initial application of representing soft set in Boolean information system for decision making (P. Maji et al., 2002). Representation of soft set in Boolean information system is already discussed in Example 1.1.

2.1.1.1 Obtaining reduct table and decision making

PK Maji approach calculates all reduct sets first. Then the choice values c_i for reduct soft set is calculated by summing up all values for each object using below relation.

$$c_i = \sum_j h_{ij} \qquad (2.1)$$

The maximum choice value c_k of any reduct set is selected as the optimal choice as explained in below example

Example 2.1: Reduct and decision making in Soft Set using PK Maji approach

Suppose Mr. X is interested in buying house on the bases of parameter having subset P ={beautiful, wooden, cheap, in green surrounding, in good repair} $=\{e_1,e_2,e_3,e_4,e_5\}$. Then the tabular representation for (F,P) is given in Table 2.1.

Table 2.1: Representation of (F,P), for finding Mr. X choice

$U\|P$	e_1	e_2	e_3	e_4	e_5	c_i
h_1	1	1	1	1	1	5
h_2	1	1	1	1	0	4
h_3	1	0	1	1	1	4
h_4	1	0	1	1	0	3
h_5	1	0	1	0	0	2
h_6	1	1	1	1	1	5

According to PK Maji, the sub sets $(F,Q)=\{e_1,e_2,e_4,e_5\}$ and $(F,R)=\{e_1,e_3,e_4,e_5\}$ are two reduct soft sets of soft set (F,P). Any of them can be selected for calculating choice of Mr. X. Let the sub set (F,Q) is chosen as reduct with its choice values c_i as given in Table 2.2.

Table 2.2: PK Maji Reduct soft Set (F,Q) of (F,P)

$U\|Q$	e_1	e_2	e_4	e_5	c_i
h_1	1	1	1	1	4
h_2	1	1	1	0	3
h_3	1	0	1	1	3
h_4	1	0	1	0	2
h_5	1	0	0	0	1
h_6	1	1	1	1	4

It can be observed from Table 2.2 that h_1 and h_6 have highest c_i value, therefore either of them is best choice or optimal choice for Mr. X.

2.1.2 The Parameterization reduction

D Chen et al. pointed out that the approach of getting reduct table by PK Maji is incorrect. Decision or choice value must be calculated before reduct (Degang Chen et al., 2005). Furthermore, they extended the concept of rough set parameter reduction (Peng, Kolda, & Pinar, 2014) to obtain reduct in soft set. Before reviewing Chen approach, few important definitions are presented below.

Let U is a set of objects and (F,A) and (G,B) are two soft sets over U. Let * denote a binary operation.

Definition 2.1: $(F,A)*(G,B)=(H,\ A\times B)$, where $H(\alpha,\beta)=F(\alpha)*G(\beta)$, $\alpha\in A$, $\beta\in B$ and A×B is the Cartesian product of set A and B.

Definition 2.2: if $B\subseteq A$ then a binary relation called indiscernibility denoted by $IND(B)$ and given by

$$IND(B) = \{(x, y) \in U \times U : a(x) = a(y) \forall a \in B\}$$

In other words, indiscernibility is an equivalence relation given by

$$IND(B) = \bigcap_{\alpha \in B} IND(\alpha)$$

Definition 2.3: Suppose *R* is the family of equivalence relations and let $A \subseteq R$. *A* is said to be dispensable in *R* if $IND(R) = INR(R - A)$. If *A* is dispensable in *R* then $R - A$ is a reduct of *R*.

Consider Example 2.1, choice values for all objects are calculated using first D Chen approach in Table 2.3. Mr. X choice is maximum of c_i which is $h_1 = h_6 = 5$. So, Mr. X can choose any of these houses as an optimal choice.

Table 2.3: Choice values calculation for Mr. X using D Chen approach

$U\|P$	e_1	e_2	e_3	e_4	e_5	c_i
h_1	1	1	1	1	1	5
h_2	1	1	1	1	0	4
h_3	1	0	1	1	1	4
h_4	1	0	1	1	0	3
h_5	1	0	1	0	0	2
h_6	1	1	1	1	1	5

According to Definition 2.3, if e_1 and e_3 are deleted from the table, there will be no effect on Mr. X choice and it remains same. Therefore, $\{e_1, e_3\}$ is dispensable in *P* and $P - \{e_1, e_3\}$ is the reduct set of *P* as given in Table 2.4.

Table 2.4: D Chen Reduct for Mr. X Choice

U|*(P-R)*	e_2	e_4	e_5	c_i
h_1	1	1	1	3
h_2	1	1	0	2
h_3	0	1	1	2
h_4	0	1	0	1
h_5	0	0	0	0
h_6	1	1	1	3

It can be observed form Table 2.4, that optimal choice for Mr. X is still h_1 and h_6 because both have maximum choice values in the reduct table as well.

2.1.3 Normal Parameter Reduction

This method presented by Z Kong discloses below two issues in parameterization reduction technique of D Chen.

2.1.3.1 Flaws of Parameterization Reduction

First problem of D Chen approach is that, the reduct calculated is not valid for getting sub-optimal choices. Secondly, if a set of new attributes is added to both original and its Chen reduct table, the choices of new resulted tables is different from original and reduct tables. These problems are explained in Example 2.2 taken from Z Kong article (Kong et al., 2008).

Example 2.2: consider Table 2.5 is an original soft set. Parameterization reduction of original table is given in Table 2.6 and h_2 is the optimal choice for both original and its reduct table. A new table of parameters e_1^*, e_2^* and e_3^* is added into both original table and its reduct table as given in Table 2.7 and 2.8 respectively. In both new tables, the optimal choice is changed from h_2 to h_1 and h_3. It can also be observed from original table and its reduct table that original sub optimal choice are h_1 and h_6 while it is changed to all objects except optimal in reduct table.

Table 2.5: Original soft set example

U|E	e_1	e_2	e_3	e_4	e_5	e_6	e_7	c_i
h_1	1	0	1	1	1	0	0	4
h_2	0	0	1	1	1	1	1	5
h_3	0	0	0	0	0	1	1	2
h_4	1	0	1	0	0	0	0	2
h_5	1	0	1	0	0	0	0	2
h_6	0	1	1	1	0	1	0	4

Table 2.6: Reduct table of original table

U|R	e_3	e_6	c_i
h_1	1	0	1
h_2	1	1	2
h_3	0	1	1
h_4	1	0	1
h_5	1	0	1
h_6	1	1	1

Table 2.7: Original table combined with new parameters

U|E+*	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e^*	e^*	e^*	c_i
h_1	1	0	1	1	1	0	0	1	0	1	6
h_2	0	0	1	1	1	1	1	0	0	0	5
h_3	0	0	0	0	0	1	1	1	1	1	5
h_4	1	0	1	0	0	0	0	0	0	1	3
h_5	1	0	1	0	0	0	0	1	1	0	4
h_6	0	1	1	1	0	1	0	1	0	0	5

Table 2.8: Reduct table combined with new parameters

U|R+*	e_3	e_6	e^*_1	e^*_2	e^*_3	c_i
h_1	1	0	1	0	1	3
h_2	1	1	0	0	0	2
h_3	0	1	1	1	1	4
h_4	1	0	0	0	1	2
h_5	1	0	1	1	0	3
h_6	1	1	1	0	0	3

It is clear from Example 2.2 that D Chen technique of parameterization reduction is not applicable to sub-optimal choices calculation and optimal choice calculation is inconsistent in added parameters.

2.1.3.2 Normal parameters reduction and Solution to the flaws of Parameterization reduction

Kong et al. presented Normal parameter reduction as a solution to the above problems of parameterization reduction (Kong et al., 2008). They presented an algorithm for their technique that uses a lot of mathematics and details can be found in their related article. Here, without going to algorithmic description and mathematical details, their approach is briefly explained with a necessary definition and example.

Definition 2.4: if there exists $A \subset E$ for a soft set (F, E) such that $c_{A1} = c_{A1} = c_{A3} = \cdots = c_{An}$ then *A* is dispensable in E and $E - A$ is reduct set of soft set (F, E). Where, $c_{A1}, c_{A1}, c_{A3}, \cdots, c_{An}$ are the choice values of parameter set *A* for object 1 to *n*.

Example 2.3: Consider the soft set of Table 2.5, $A = \{e_1, e_2, e_7\} \subset E$. According to Definition 2.4, all c_{Ai} have same values equal to 1 as given in Table 2.9. Therefore, *A* is dispensable in *E*, $E - A = \{e_3, e_4, e_5, e_6\}$ is the reduct set of *E* and $(F, E - A)$ is the reduct soft set of (F, E) as given in Table 2.10.

Table 2.9: Dispensable set A in E

U|A	e_1	e_2	e_7	c_i
h_1	1	0	0	1
h_2	0	0	1	1
h_3	0	0	1	1
h_4	1	0	0	1
h_5	1	0	0	1
h_6	0	1	0	1

Table 2.10: Normal Parameter reduction of original table

U|E-A	e_3	e_4	e_5	e_6	c_i
h_1	1	1	1	0	3
h_2	1	1	1	1	4
h_3	0	0	0	1	1
h_4	1	0	0	0	1
h_5	1	0	0	0	1
h_6	1	1	0	1	3

It can be observed from original Table 2.5 and its reduct Table 2.10 that optimal as well as sub optimal choices are same while obtained by normal parameters reduction. New parameters $e^*{}_1, e^*{}_2$ and $e^*{}_3$ are added to reduct Table 2.10 in Table 2.11 to check its consistency with original table for optimal and sub optimal choice. It can be observed from added parameters original Table 2.7 and added parameters redcut Table 2.11 that in both tables h_1 is the optimal choice, h_2, h_3 and h_6 are the first sub-optimal choices, h_5 is the second sub-optimal choice and h_4 is the last sub-optimal choice.

Table 2.11: Added parameters to Normal parameters reduction table

U|E+*	e_3	e_4	e_5	e_6	$e^*{}_1$	$e^*{}_2$	$e^*{}_3$	c_i
h_1	1	1	1	0	1	0	1	5
h_2	1	1	1	1	0	0	0	4
h_3	0	0	0	1	1	1	1	4
h_4	1	0	0	0	0	0	1	2
h_5	1	0	0	0	1	1	0	3
h_6	1	1	0	1	1	0	0	4

Hence, normal parameters reduction has the consistency in optimal to sub-optimal choices in adding new parameters.

2.1.4 New Efficient Normal Parameters Reduction

New efficient normal parameters reduction technique is presented to overcome the computational complexity and much mathematical involvement in previous approach of Kong's normal parameter reduction. In this technique, reduct table has same consistency with original table in calculating optimal to sub-optimal choices and adding new parameters. But the algorithm of this technique is easy to understand, short and has less computational complexity (Ma, Sulaiman, Qin, Herawan, & Zain, 2011).

2.2 Incomplete Soft set and Its Handling Techniques

In the previous section, the major applications of soft set were discussed. Most probable reasons are mentioned in the upcoming section, due to which a soft set might get some values missing. If a soft set contains missing values due to any reason, it becomes incomplete soft set. Incomplete soft set can no longer be used in these applications and if still used will result in misleading results. In this section, existing techniques of dealing with incomplete soft set are discussed in detail.

2.2.1 Reasons of incompleteness in soft set

Data of soft set can be missed due to any of the following reasons.

i. **Human mistakes**: humans can miss, exclude or ignore some values during data entry. This mistake can be both intentional or unintentional

ii. **Machine errors:** data can be missed from machine too after its proper entry by humans. This can be caused by some interrupt like power failure or hardware malfunctioning.

iii. **Virus attacks:** malwares and viruses can also alter the arrangements of data after proper saving

iv. **Security reasons:** sometimes all data can't be entered or transferred due to security and privacy reasons.

v. **Communicational errors:** data can be missed due to reasons like loss in signals and dispersions during transferring it from one point to another through communication mediums.

2.2.2 Incomplete Soft Set

An information system $S^* = (U, AT, V_r, f)$ is called incomplete if $f(x_i, a_j)$ is not known, where, $U = (x_1, x_2, \cdots, x_n)$, $AT = (a_1, a_2, \cdots, a_m)$, $x_i \in U$, $i = 1,2,3,\cdots, n$ and $a_j \in AT$ information system, where unknown entries in the table are represented by symbol "*" for $j = 1,2,3,\cdots, m$. The following example presents an incomplete soft set.

Example 2.4: Suppose $U = (s_1, s_2, s_3, \cdots, s_8)$ is a set of applicants with parameters set $E = \{e_1, e_2, e_3, e_4, e_{5\,,} e_6\}$ representing "young age", "experienced", "married", "the highest academic degree is Master", "studied abroad", and "the highest academic degree is Doctor", respectively with its soft set illustration in presented as in Table 2.12.

Table 2.12: Representation of incomplete soft set

U/E	e_1	e_2	e_3	e_4	e_5	e_6
s_1	0	1	1	1	0	0
s_2	0	1	0	0	0	1
s_3	1	0	0	1	0	0
s_4	1	0	$*_1$	0	$*_2$	1
s_5	0	1	1	0	0	1
s_6	1	0	0	$*_3$	0	0
s_7	$*_4$	1	1	1	0	0
s_8	0	0	1	0	0	1

From incomplete Table 3.12, it is known that candidate 4 is young, inexperienced, having Ph.D. as his highest degree, but it is unknown that whether he is married and studied abroad or not. Similarly, for candidate 6 and 7, the “highest degree is master” and “young age” values are unknown respectively. Hence it is an incomplete soft set with unknown values represented by $*_1, *_2, *_3$ and $*_4$.

2.2.3 Data Analysis Approaches

Data analysis approaches of soft set under incomplete information uses weighted average technique for decision value calculation of incomplete soft set while incomplete data in fuzzy soft set is predicted through average probability (Zou & Xiao, 2008). Here, in relation to proposed work, their soft set case is discussed only. According to this approach the decision value d_i among all objects is calculated using below relation

$$d_i = \sum_{i=1}^{m} k_i c_i \tag{2.2}$$

where c_i is the choice value of each object, *m* is maximum number of choices for same object having missing value and k_i is the weight of choice values. For one missing value, the choice values of an object are only two (0 or 1) and its respected weights are $k_1 = \frac{n_0}{n_0 + n_1} = q_{e_i}$ and $k_2 = \frac{n_1}{n_1 + n_0} = p_{e_i}$. For more than one missing values *t* of same object, the choice values increase and its respective weight values are calculated by

$$k = \begin{cases} \prod_{e \in E_0^*} q_e & x = 0, \\ \sum_{C_x^t} \left(\left(\prod_{e_i \in E_1^*} p_{e_i} \right) \left(\prod_{e_j \in E_0^*} q_{e_j} \right) \right) & 0 < x < t, \\ \prod_{e \in E_1^*} p_e & x = t \end{cases}$$

Where, x is the number of 1s in the row, while E_1^* and E_0^* are its parameter sets for value 1 and 0 respectively. Using this approach, the decision value in term of candidate's eligibility for incomplete Table 2.12 is calculated as explained in the related article (Zou & Xiao, 2008) and given in Table 2.13.

Table 2.13: Decision value calculated by Zou *et al.* technique for incomplete soft set of Example 2.4

U/E	e_1	e_2	e_3	e_4	e_5	e_6	d_i
s_1	0	1	1	1	0	0	3
s_2	0	1	0	0	0	1	2
s_3	1	0	0	1	0	0	2
s_4	1	0	$*_1$	0	$*_2$	1	2.57
s_5	0	1	1	0	0	1	3
s_6	1	0	0	$*_3$	0	0	1.43
s_7	$*_4$	1	1	1	0	0	3.43
s_8	0	0	1	0	0	1	2

2.2.4 Using Parity Bits and Supported Set

In this approach, soft set is represented in Boolean valued information system. Supported sets from all objects and even parity bits for each row and column are calculated for a completed table (having no missing information at the time of these calculations). Few missed values can be recalculated using available supported sets and parity bits values. (Rose et al., 2011)

2.2.4.1 Supported Set

It is simply the arithmetic sum of values of an object or number of 1s in a row. Mathematically for object *u*

$$\operatorname{supp}(u) = \operatorname{card}\left(e \in E{:}f\,(u,e) = 1\right) \tag{2.3}$$

And the set of $\operatorname{supp}(u)$ for all objects is supported set.

2.2.4.2 Even parity bits for rows and columns

A bit column is put for making the bit’s parity of each object even. 0 is put in parity bit column if object has already even number of 1s, otherwise, 1 is put. Mathematically for object *u*

$$P_{bit} = \operatorname{supp}(u) \mod 2 \tag{2.4}$$

Similarly, for an attribute or column, the parity bit is defined as

$$C_{bit} = \left(\sum_{i=1}^{n} f\,(u,e_i)\right) \mod 2 \tag{2.5}$$

Their technique is explained in Example 2.5.

Example 2.5: Consider a soft set (F,E) representing the communication skill of university students. For ten students, $U = \{u_1,\, u_2,\, u_3,\, u_4,\, u_5,\, u_6,\, u_7,\, u_8,\, u_9,\, u_{10}\}$ and $E = \{e_1,\, e_2,\, e_3,\, e_4,\, e_5,\, e_6\}$ whose parameters stands for using communication facilities as email, Facebook, blog, Friendster, yahoo messenger and SMS respectively. (F,E) is represented according to its approximation Table 2.14.

Table 2.14: Representation of Soft Set (F, E) for Example 2.5

U/E	e_1	e_2	e_3	e_4	e_5	e_6
u_1	1	0	1	0	0	0
u_2	0	1	1	1	1	1
u_3	1	0	0	1	1	1
u_4	1	1	1	1	0	0
u_5	0	1	0	0	0	0
u_6	0	0	1	0	0	0
u_7	0	0	0	0	1	0
u_8	1	1	1	1	1	0
u_9	1	1	1	0	1	1
u_{10}	1	0	0	1	0	0

Support sets and parity bits' values for objects, parameter parity bits for Example 2.5 of Table 2.14 are calculated in Table 2.15 as following.

Table 2.15: Supported Set and Parity Bit Calculation for (F, E) of Example 2.5

U/E	e_1	e_2	e_3	e_4	e_5	e_6	P_{bit}	Supp
u_1	1	0	1	0	0	0	0	2
u_2	0	1	1	1	1	1	1	5
u_3	1	0	0	1	1	1	0	4
u_4	1	1	1	1	0	0	0	4
u_5	0	1	0	0	0	0	1	1
u_6	0	0	1	0	0	0	1	1
u_7	0	0	0	0	1	0	1	1
u_8	1	1	1	1	1	0	1	5
u_9	1	1	1	0	1	1	1	5
u_{10}	1	0	0	1	0	0	0	2
C_{bit}	0	1	0	1	1	1		--

After having these calculations, suppose that few values i.e. u_{13}, u_{22}, u_{24}, u_{33},u_{34}, u_{35}, u_{41}, u_{44}, u_{45}, u_{54} and u_{65} are missing, as shown by *s in Table 2.16.

Table 2.16: Missing values Representation

U/E	e_1	e_2	e_3	e_4	e_5	e_6	P_{bit}	Supp
u_1	1	0	*	0	0	0	0	2
u_2	0	*	1	*	1	1	1	5
u_3	1	0	*	*	*	1	0	4
u_4	*	1	1	*	*	0	0	4
u_5	0	1	0	*	0	0	1	1
u_6	0	0	1	0	*	0	1	1
u_7	0	0	0	0	1	0	1	1
u_8	1	1	1	1	1	0	1	5
u_9	1	1	1	0	1	1	1	5
u_{10}	1	0	0	1	0	0	0	2
C_{bit}	0	1	0	1	1	1	--	

For u_{13} it can be noticed that $P_{bit}(u_1) = 0$, so u_{13} can be put as 1 easily i.e. $u_{13} = 1$. For u_{22} the $C_{bit}(e_2) = 1$, so $u_{22} = 1$. Similarly, in remaining row/columns, single missing values $u_{41} = 1$and $u_{65} = 0$. The missing values reduce to Table 2.17.

Table 2.17: Calculating single missing values in a column or row using parity bit

U/E	e_1	e_2	e_3	e_4	e_5	e_6	P_{bit}	Supp
u_1	1	0	1	0	0	0	0	2
u_2	0	1	1	*	1	1	1	5
u_3	1	0	*	*	*	1	0	4
u_4	1	1	1	*	*	0	0	4
u_5	0	1	0	*	0	0	1	1
u_6	0	0	1	0	0	0	1	1
u_7	0	0	0	0	1	0	1	1
u_8	1	1	1	1	1	0	1	5
u_9	1	1	1	0	1	1	1	5
u_{10}	1	0	0	1	0	0	0	2
C_{bit}	0	1	0	1	1	1	--	

In Table 2.17, the values u_{24}, u_{33} and u_{54} become single which take the values 1, 0 and 0 respectively. Another Table 2.18 is obtained below after putting these values.

Table 2.18: Calculating consecutive two missing values in a column or row using parity bit and supported set

U/E	e_1	e_2	e_3	e_4	e_5	e_6	P_{bit}	Supp
u_1	1	0	1	0	0	0	0	2
u_2	0	1	1	1	1	1	1	5
u_3	1	0	0	*	*	1	0	4
u_4	1	1	1	*	*	0	0	4
u_5	0	1	0	0	0	0	1	1
u_6	0	0	1	0	0	0	1	1
u_7	0	0	0	0	1	0	1	1
u_8	1	1	1	1	1	0	1	5
u_9	1	1	1	0	1	1	1	5
u_{10}	1	0	0	1	0	0	0	2
C_{bit}	0	1	0	1	1	1	--	

In Table 2.18, object u_3 has two missing values, since its parity bit is 0 and support value is 4 so $u_{34} = u_{35} = 1$. In u_{44} and u_{45} the parity bit is 0, which means that either of them is 1. Form C_{bit} it can calculated that $u_{44} = 1$, therefore $u_{45} = 0$. Hence, a complete Table 2.19 is obtained which is same as Table 2.14.

Table 2.19: Complete Soft set after calculating all missing values

U/E	e_1	e_2	e_3	e_4	e_5	e_6
u_1	1	0	1	0	0	0
u_2	0	1	1	1	1	1
u_3	1	0	0	1	1	1
u_4	1	1	1	1	0	0
u_5	0	1	0	0	0	0
u_6	0	0	1	0	0	0
u_7	0	0	0	0	1	0
u_8	1	1	1	1	1	0
u_9	1	1	1	0	1	1
u_{10}	1	0	0	1	0	0

2.2.5 Using rows, columns and diagonals aggregates

This approach is an extended from of previous one. In addition to rows and columns support values, the aggregate values of diagonals are also used for calculating missing data (Rose et al., 2011). Its performance is improved to calculate more consecutive missing values. Their algorithm is given in Figure 1 and the technique is explained it with example

2.2.5.1 Attribute aggregate values

It is the arithmetic sum of an attribute values

$$C_{agg} = \sum_{i=1}^{n} f\left(u_i, e\right) \tag{2.6}$$

2.2.5.2 Diagonal aggregate values

For a table representing soft set having u_i objects and parameter set *E*, a tuple or diagonal can be expressed mathematically

$$t_i = \left(f\left(u_i, a_1\right),\ f\left(u_i, a_2\right),\ f\left(u_i, a_3\right),\ \ldots,\ f\left(u_i, a_{|A|}\right)\right) \tag{2.7}$$

Where, $i = 1,\ 2,\ 3, \ldots, |U|$

If *D* is the number of unidirectional diagonals in a table, then

$$D = |U| + |A| \text{ -1} \tag{2.8}$$

As rows and column are treated horizontally and vertically, likewise, diagonals can be dealt in left to right (LR) and right to left (RL) manners for getting two dimensional accumulate values. As it is noticed that number of diagonals (D) is more than the number of column or rows, therefore both LR and RL diagonals have two different cases.

Case 1: For $1 \leq k \leq |A|$

$$Diag_{LR}(k) = \sum_{i=1}^{k} f\left(u_i,\, a_j\right) \qquad (2.9)$$

Where, $j = k - i + 1$

$$Diag_{RL}(k) = \sum_{i=1}^{k} f\left(u_i,\, a_j\right) \qquad (2.10)$$

Where, $j = |A| - k + i$

Case 2: For $|A| < k < D$

$$Diag_{RL_u}(k) = \sum_{j=k-|A|+1}^{|U|} f\left(u_i,\, a_j\right) \qquad (2.11)$$

Where, $j = k - i + 1$, for $i \leq k$ and $j \leq |U|$

$$Diag_{LR_u}(k) = \sum_{j=k-|A|+1}^{|U|} f\left(u_i,\, a_j\right) \qquad (2.12)$$

Where, $j = |A| - k + i$, for $i \leq k$ and $j \leq |U|$

```
Calculating missing values from aggregates

  Input: Partially incomplete Boolean information table and
aggregate values
  Output: Complete Boolean information table

1. Calculate supported values of rows, aggregate values of
   columns and diagonals.
2. Find every single value first by applying horizontal or
   vertical or diagonal summation
3. Repeat step 2 until no single value remains
4. Find other missing values applying supported and or column
   aggregate and or diagonal aggregate.
```

Figure 2.1: Calculating partial missing values from aggregates

Example 2.6: Consider the complete soft set as given in Table 2.20, its rows and columns, LR and RL diagonal aggregates values are calculated in Table 2.21, to Table 2.23.

Table 2.20: A complete soft set representation in tabular form

U/E	e_1	e_2	e_3	e_4	e_5
h_1	0	1	0	1	1
h_2	1	0	0	0	0
h_3	0	1	1	1	0
h_4	1	0	1	0	0
h_5	0	0	1	1	0
h_6	0	0	0	0	0

Table 2.21: Rows and columns aggregate values

U/E	e_1	e_2	e_3	e_4	e_5	Supp(h)
h_1	0	1	0	1	1	3
h_2	1	0	0	0	0	1
h_3	0	1	1	1	0	3
h_4	1	0	1	0	0	2
h_5	0	0	1	1	0	2
h_6	0	0	0	0	0	0
$\sum col$	2	2	3	3	1	

Table 2.22: Left to Right (LR) aggregates

U/E	e_1	e_2	e_3	e_4	e_5	Diagonal aggregate	
h_1	0	1	0	1	1		
h_2	1	0	0	0	0	1	$Diag_{LR}(k)$
h_3	0	1	1	1	0	1	
h_4	1	0	1	0	0	0	
h_5	0	0	1	1	0	2	
h_6	0	0	0	0	0	1	
						4	$Diag_{LR_u}(k)$
						1	
						1	
						0	
						0	

Table 2.23: Right to Left (RL) aggregates

Diagonal aggregate		e_1	e_2	e_3	e_4	e_5	U/E
		0	1	0	1	1	h_1
$Diag_{RL}(k)$	0	1	0	0	0	0	h_2
	2	0	1	1	1	0	h_3
	0	1	0	1	0	0	h_4
	3	0	0	1	1	0	h_5
	2	0	0	0	0	0	h_6
$Diag_{RL_u}(k)$	2						
	2						
	1						
	0						
	0						

Now suppose there are some values missing in this example as shown in Table 2.24.

Table 2.24: Soft set with supposed missing values

U/E	e_1	e_2	e_3	e_4	e_5
h_1	*	1	0	1	1
h_2	1	0	0	0	0
h_3	0	1	1	1	0
h_4	1	*	*	0	0
h_5	0	*	*	*	0
h_6	0	0	0	0	0

In Table 2.24, missing values are $u_{11}, u_{42}, u_{43}, u_{52}, u_{53}$ and u_{54}. From Table 2.21, column aggregate of e_1 is 2, therefore $u_{11} = 0$. Similarly, $u_{54} = 0$. For u_{42}, RL equal to 2 from Table 2.22, therefore, $u_{42} = 0$. Similarly, all other missing values can be found easily.

2.2.6 Novel Data Filling Approach for an Incomplete Soft Set (DFIS)

The approach proposed by Qin *et al.* prefers to predict missing value through association between parameters. This association is considered as the first case of their approach (Qin, Ma, Herawan, & Zain, 2011). For instance, in Example 1.1, it is inconsistent association that a house in good repair can’t be in bad repair, cheap can’t be expensive. Similarly, in same example beautiful houses and houses in good repair are most probably expensive is consistent association. In Example 2.1, a highest degree can be either master or doctorial and young age candidate is more probably inexperienced and unmarried, indicating inconsistent associations. Similarly, more consistent and inconsistent associations can be found between parameters. Mathematical description of this technique is explained below.

The consistent association between two parameters is found by

$$CN_{ij} = \left| \left\{ x \mid F_{e_i}(x) = F_{e_j}(x),\ x \in U_{ij} \right\} \right| \qquad (2.13)$$

Where CN_{ij} is the number of elements in column (parameter) *i* having same value to the number of parameter (column) *j*.

Consistent association degree is calculated by

$$CD_{ij} = \frac{CN_{ij}}{\left|U_{ij}\right|} \tag{2.14}$$

Where $\left|U_{ij}\right|$ is the cardinality (absolute number) of known element's pairs for parameter *i* and *j*. i.e. CD_{ij} is the ratio of consistency to number of total elements in columns *i* and *j*.

Similarly, inconsistent association is found as

$$IN_{ij} = \left| \left\{ x \left| F_{e_i}(x) \neq F_{e_j}(x), \ x \in U_{ij} \right. \right\} \right| \tag{2.15}$$

Inconsistent association degree is calculated by

$$ID_{ij} = \frac{IN_{ij}}{\left|U_{ij}\right|} \tag{2.16}$$

To know that whether the association is consistent or inconsistent, net association degree is obtained by

$$D_{ij} = \max\left\{CD_{ij}, ID_{ij}\right\} \tag{2.17}$$

To find the two parameters having maximum association with each other, the maximal association degree is got among the set of all association degrees by

$$D_i = \max\left\{D_{ij}\right\} \tag{2.18}$$

As a result, the unknown(s) value $F_{e_i}(x)$ is predicted as same as the corresponding element(s) *j* (0 for 0 and 1 for 1) if the association is consistent, otherwise it is predicted as a complement of the parameter *j* for inconsistent association.

In second case, when there is weak association between parameters i.e. $|D_i| < \lambda$, where λ is a pre-set threshold value. Then, probability for zero and one is calculated as $p_1 = \dfrac{n_1}{n_1 + n_0}$ and $p_0 = \dfrac{n_0}{n_0 + n_1}$, where n_1 and n_0 are the number of 1s and 0s respectively for the parameter having missing data. As a result, the missing value is put as 1 if $p_1 > p_o$, 0 if $p_1 < p_o$ and either 1 or 0 if $p_1 = p_o$. The following Example explains DFIS approach step by step.

Example 2.7: Predicting values through DFIS for incomplete case of Example 2.4. Here the parameters e_1, e_3, e_4 and e_5 have missing data.

Step 1: Finding consistency CN_{ij} and inconsistency IN_{ij}

Parameter 1 with 2: as only s_8 has the same value equal to 0 for both e_1 and e_2, therefore, $CN_{12} = 1$, as the values are not same for all other 6 objects excluding the missing s_7, therefore, $IN_{12} = 6$. Similarly, ($CN_{13} = 1$, $IN_{13} = 5$), ($CN_{14} = 4$, $IN_{14} = 2$), ($CN_{15} = 4$, $IN_{15} = 2$) and ($CN_{16} = 2$, $IN_{16} = 5$).

Step 2: Calculating ratio of consistency CD_{ij} and ratio of inconsistency ID_{ij}

First, finding the cardinality ($|U_{ij}|$) is needed for calculating CD_{ij} and ID_{ij}. As parameters 1 and 2 have seven complete pairs for all objects except object s_7, therefore, $|U_{12}| = 7$. Similarly, $|U_{13}| = |U_{14}| = |U_{15}| = 6$ and $|U_{16}| = 7$.

Hence, CD_{12} = $CN_{12} \big/ \left|U_{12}\right|$ = 1/7 = 0.14 and $ID_{12} = 0.86$. Similarly, ($CD_{13} = 0.16$, $ID_{13} = 0.83$), ($CD_{14} = 0.67$, $ID_{14} = 0.33$), ($CD_{15} = 0.67$, $ID_{15} = 0.33$) and ($CD_{16} = 0.28$ $ID_{16} = 0.83$).

Step 3: Deciding whether association is consistent or inconsistent

As D_{ij} = $\max\{CD_{ij}, ID_{ij}\}$, therefore, $D_{12} = \max\{CD_{12}, ID_{12}\} = \max\{0.86, 0.14\}$=0.86. As the association is inconsistent therefore, minus (-) sign will be used for its indication and differentiation from consistent one i.e. $D_{12} = -0.86$. Similarly, $D_{13} = -0.83$, $D_{14} = 0.67$, $D_{15} = 0.67$ and $D_{16} = -0.83$.

Step 4: Calculating maximal degree of association

D_{ij} is calculated according to step 3 for those parameters having missing values e_1, e_3, e_4 and e_5 with all other parameters e_1, e_2, e_3, … , e_6 as presented in Table 2.25.

Table 2.25: Calculation of D_{ij} for incomplete Table 2.12

E^* / E	e_1	e_2	e_3	e_4	e_5	e_6
e_1	--	-0.86	-0.83	0.67	0.67	-0.83
e_3	-0.83	0.71	--	±0.5	-0.67	0.57
e_4	0.67	0.57	±0.5	--	±0.5	-1
e_5	0.67	-0.57	0.57	±0.5	--	0.57

From Table 2.25, it can be seen that for e_1, $D_1 = \max\{D_{12}, D_{13}, D_{14}, D_{15}, D_{16}\}$= $\max\{0.86, 0.83, 0.67, 0.67, 0.83\}$ = -0.86. Similarly, $D_3 = -0.83$, $D_4 = -1$ and $D_5 = 0.67$.

Step 5: Putting values according to association

The threshold is set to 0.85 i.e. λ=0.85. Only e_1 and e_4 are satisfying the condition to be calculated by association because, $D_1 = |-0.86| > \lambda$ and $D_4 = |-1| > \lambda$. From Table 2.25, e_1 has inconsistent association with e_2 and the corresponding element (u_{72}) of its missing element ($*_4 = u_{71}$) has the value equal to 1 in Table 2.12. As complement value is assigned in case of inconsistent association, therefore, $*_4$=0. Similarly, $*_3$=1.

Step 6: Calculating probabilities for weak association.

As D_3 and D_5 have smaller values than the fixed threshold λ=0.85. Therefore, $*_1$ and $*_2$ can't be calculated through association, rather probability will be used for predicting these values. For e_3 it can be seen that $n_1 = 4$ and $n_0 = 3$ implies that $p_1 = \frac{4}{4+3} = 0.57$ and $p_0 = \frac{3}{3+4} = 0.43$, as $p_1 > p_0$, therefore, $*_1 = 1$. Similarly, $*_2 = 0$. A complete Table 2.26 is obtained after putting these predicted values in incomplete Table 2.12.

Table 2.26: Incomplete Soft Set Completed Using DFIS

U/E	e_1	e_2	e_3	e_4	e_5	e_6
s_1	0	1	1	1	0	0
s_2	0	1	0	0	0	1
s_3	1	0	0	1	0	0
s_4	1	0	1	0	0	1
s_5	0	1	1	0	0	1
s_6	1	0	0	1	0	0
s_7	0	1	1	1	0	0
s_8	0	0	1	0	0	1

2.2.7 An efficient decision making approach in incomplete soft set

The approach proposed by Kong *et al.* (Kong et al., 2014) is equivalent to Zou *et al.* approach (Zou & Xiao, 2008) in results but more simplified with respect to complexity. Instead of using weighted-average huge computations, its uses simple probability $p'_{e_j} = \frac{n_1}{n_1 + n_0}$ for calculating an unknown value, where n_1 and n_0 are the number of 1 and 0 respectively for same parameter. After inserting this value in unknown the decision value is calculated by $d_i = \sum_{j=1}^{m} h_{ij}$. Using this technique, the incomplete Example 2.4 gets completed as given in Table 2.27 along with decision value d_i.

Table 2.27: Incomplete soft set Table 2.12 after completion and d_i calculation using Kong approach

U/E	e_1	e_2	e_3	e_4	e_5	e_6	d_i
s_1	0	1	1	1	0	0	3
s_2	0	1	0	0	0	1	2
s_3	1	0	0	1	0	0	2
s_4	1	0	$\frac{4}{4+3}$	0	$\frac{0}{0+7}$	1	2.57
s_5	0	1	1	0	0	1	3
s_6	1	0	0	$\frac{3}{3+4}$	0	0	1.43
s_7	$\frac{3}{3+4}$	1	1	1	0	0	3.43
s_8	0	0	1	0	0	1	2

2.3 Link prediction and community detection in OSNs

The literature of this part is divided into three parts; the first part contains previous prominent techniques that worked on link prediction and network community detection, the second part consists on ranking algorithms, while finding spreading efficiencies of network nodes for evaluating the proposed prediction method is described in third part.

2.3.1 Link prediction

Researchers have attempted to detect network communities (Bedi & Sharma, 2016; Fortunato, 2010; Palla, Barabási, & Vicsek, 2007; Peng et al., 2014; Radicchi et al., 2004; Sun, 2016; Zhan, Guan, Chen, Niu, & Jin, 2016), proposed various definitions, and concluded that “its elements are highly interconnected” (Güneş, Gündüz-Öğüdücü, & Çataltepe, 2016). Progress has been achieved in terms of completing an incomplete network (i.e., an OSN) by predicting new links (Adamic & Adar, 2003; Duan, Aggarwal, Ma, Hu, & Huai, 2016; Güneş et al., 2016; Kossinets, 2006; D. Li, Zhang, Xu, Chu, & Li, 2016; Liben-Nowell & Kleinberg, 2007; Lü & Zhou, 2011; Newman, 2001). Link prediction is divided into two categories: network topology based and node based (Güneş et al., 2016). Link prediction approaches that use network topology are based on the fact that communities utilize different aspects of common neighbors but their main focus is on “interconnection among nodes” with its own significance (Güneş et al., 2016; Zhan et al., 2016).

2.3.2 Ranking Algorithms

Researchers have proposed various algorithms to detect and rank top spreaders in OSNs. Among these, PageRank and *k*-core are considered the most outstanding and widely used algorithms.

2.3.2.1 PageRank

PageRank is a network-based diffusion algorithm originally proposed by Brin et al. (Brin & Page, 2012). This well-known algorithm is used by the Google search engine for ranking web pages. It allows for the global ranking of all web pages based only on their connected links and locations in the web graph, regardless of their content. PageRank calculates recursively and considers two main parameters, namely, the number of inbound links and their corresponding PageRank values.

2.3.2.2 *k*-Core ranking

In *k*-core-based ranking, each node is assigned a *k*-shell number k_s, which is the order of the shell to which it belongs. Initially, the *k*-shell eliminates all the nodes with a degree (*k*) of 1. The elimination process continues until all the nodes with a degree of 1 are eliminated. Similarly, this elimination procedure is applied to the next *k*-shells. This decomposition process is repeated until the *k*-core of the network is detected (Batagelj & Zaversnik, 2003).

2.3.3 Spreading efficiency

To evaluate the validity of the proposed link prediction method for OSN completion, this research uses the ranking algorithms PageRank and *k*-Core to identify the top spreaders before and after the completion of both networks and subsequently compare the results. The spreading efficiency or influence $inf(i)$ of each user *i* is calculated as the number of users influenced by user *i* based on the wall post data set of Facebook and the retweet data set of Twitter. These influenced users are those who propagate the information of user *i*, and $inf(i)$ is obtained using breadth-first search for user *i* (S. Pei, Muchnik, Andrade Jr, Zheng, & Makse, 2014). Information spreading is in the form of sharing the wall posts of user *i* in Facebook and retweeting his or her tweets in Twitter. The retweet network serves as an illustrative network that explains how content is propagated (De Domenico, Lima, Mougel, & Musolesi, 2013). The variable $inf(i)$ is used to calculate the average spreading efficiencies M_{avg} of the set of top spreaders under consideration. Sets of top spreaders may represent the top 1%, 5%, 10%, 20%, 30%, and 50%, and their average influence levels in wall posts and retweets are considered the standard M_{avg}. Similarly, the average influence levels of the same set of top spreaders are calculated using the ranking algorithms ($M_{PR}, M_{k(s)}$) for the network. For the comparison of the accuracy rates of the ranking algorithms, the imprecision functions ε_{PR} and $\varepsilon_{k(s)}$ for PageRank and k-Core are used as proposed in (Kitsak et al., 2010) and given as

$$\varepsilon_{PR} = 1 - \frac{M_{PR}}{M_{avg}}, \tag{2.19}$$

$$\varepsilon_{k(s)} = 1 - \frac{M_{k(s)}}{M_{avg}}, \tag{2.20}$$

The lower the value of the imprecision function(ε), the more accurate the prediction, and vice versa. An ε value that is close to 0 denotes high efficiency because the selected nodes are the same as those that contribute the most to information diffusion.

CHAPTER 3: CLASSIFICATION OF INCOMPLTE SOFT SET AND CONCEPT OF ENTIRE MISSING VALUES RECALCULATION FROM AGGREGATES[3]

3.1 Introduction

Vague or uncertain data cannot be processed using conventional mathematical tools of crisp and clear data. Special models and theories, such as fuzzy set, probability, interval mathematics, rough set, grey set, and soft set, are used for the precise handling of the uncertainties in vague data to process it before use in any application and decision. In the soft set theory (Molodtsov, 1999), an application is usually based on a standard soft set with all its values represented in a binary table known as Boolean-valued information system (BIS). Ordinary arithmetic operations and processing, such as crisp data, can be performed with BIS for use in any application. BISs are mainly used for decision-making and finding optimal choices by arithmetically adding the weights of all objects, and the parameter with the maximum value is considered as the best choice (P. Maji et al., 2002). The reduct set for the soft set BIS is defined as the subset of all parameter sets that has the same decision values of optimal choice as those of the original set (Degang Chen et al., 2005). In a modified definition, a reduct set must be able to maintain the integrity of the decision values as the original set for optimal and suboptimal choices (Kong et al., 2008), and this parameterization reduction is more efficient if the method used for its calculation is easy to understand, implement, and has less computational complexity during execution (Qin et al., 2011a). Apart from these main applications of decision-making with parameterization reduction, soft set and BIS are used in several daily life applications (Feng, Jun, et al., 2010; Feng, Li, & Leoreanu-

[3] The main idea of this chapter has already been published in ISI indexed journal “IEEE Access” with the title “Concept of Entire Boolean Values Recalculation from Aggregates in the Preprocessed Category of Incomplete Soft Sets”

.

Fotea, 2010; Herawan, 2012; Herawan & Deris, 2009b; Jiang, Tang, et al., 2011; Jun et al., 2009; Jun & Park, 2008; Mamat, Herawan, & Deris, 2013; Qin, Ma, Zain, & Herawan, 2012; Rose et al., 2011; Sulaiman & Mohamad, 2013; Yuksel et al., 2013).

These applications become worthless and may yield incorrect results if several values are lost in a given BIS. Values in a soft set can be lost because of communicational errors, virus attacks, improper entry, intentional and unintentional mistakes, security, or any other probable reasons. In cases where no equivalency information of aggregates or parity bits can be found, researchers have attempted to fill and predict them from other available set of values using weighted average (Zou & Xiao, 2008), association between parameters (Qin, Ma, Herawan, & Zain, 2011b; H. W. Qin, X. Q. Ma, T. Herawan, & J. M. Zain, 2012), and probability (Kong et al., 2014) techniques. Meanwhile, the following recalculation techniques are presented from available aggregates and parity bits (Mohd Rose et al., 2011; Rose et al., 2011).

This chapter has mainly two parts. Existing techniques of incomplete soft set are classified into two categories (UP and PP) and the capability of finding entire missing values is checked for the PP category in the first part of this chapter (while UP category techniques are analyzed in next chapter). In the second part, the concept of recalculating entire missing values from aggregates is presented. This technique is extended from the previous techniques of PP category. Important definitions and algorithm for entire Boolean values recalculation are presented and the technique is explained with the help of an example as a proof of concept.

Proposed approach uses the concept of solving simultaneous linear equations for identifying unknown variables. The proposed approach bypasses the restrictions of simultaneous linear equations, such that, the number of equations must be equal to or more than unknown variables. Unlike solving simultaneous linear equations, proposed approach has the capacity to calculate more variables than that of the given number of relations. This research takes the advantage of the binary nature and limited domain of the standard soft set. This new concept can be used by researchers to develop good applications in binary-ranged data regardless of the soft set.

3.2 Analysis of Previous Techniques and their Classification

In this section, existing approaches are classified into two main categories based on their particulars and input requirement. After categorization, PP category techniques are further analyzed for finding their recalculating capability and limitation in entire missing values recalculation.

3.2.1 Incomplete soft set handling techniques

Initial attempt in calculating decision values in incomplete soft set was made using weighted average technique (Zou & Xiao, 2008), while recently, the same decision values were calculated using simple probability of 0s and 1s in an easily understandable technique and having comparatively very less computational complexity (Kong et al., 2014). The main problem of these weighted average and probability techniques are that the actual missing values still remain missed and the integrity of standard soft set will be damaged if those missing values will be recalculated back from predicted decision values and standard soft set will get converted into fuzzy soft set (Qin et al., 2011b; H. W. Qin et al., 2012). Using association between parameters avoids the problems of weighted average and probability techniques and gives second priority to probability within binary range of standard soft set (Qin et al., 2011b; H. W. Qin et al., 2012). Meanwhile, the recalculation techniques are presented which finds the missing values from available sets of aggregates and parity bits (Mohd Rose et al., 2011; Rose et al., 2011).

3.2.2 Categorization of Incomplete soft sets:

By going through the above incomplete soft set handing techniques as discussed in the literature review chapter one by one, mainly two types of them can be found. Either a technique predicts missing values and/or decision values by taking input from other available basic values or it re-calculates the missing values from other equivalent set of values. Basic values are the binary values in standard soft set (Boolean valued information system). The first type is totally dependent on basic values and completely independent from other equivalent value sets, while the later type is dependent on both i.e. basic values and available equivalent value sets. It is obvious that the sets of available equivalent values were got by certain processing of complete standard soft set. For instance, in missing values recalculation from parity bit or aggregates, the parity bit or aggregate (equivalent information) were obtained first by processing a soft set and during obtaining these sets, no single information was missing. While in the former case, no equivalent information sets are available therefore, it is considered that no processing is done on such type of incomplete soft set. Based on above arguments, all incomplete soft set techniques are classified into below two categories.

3.2.2.1 Pre-Processed Incomplete Soft set:

This category has below two previous approaches.

i. Using Parity Bits and Supported Set (Rose et al., 2011)

ii. Using rows, columns and diagonals aggregates (Mohd Rose et al., 2011)

3.2.2.2 Unprocessed[4] Incomplete Soft Set

Below three previous approaches include in this category

[4] Because of different category, we don’t further discuss UP techniques in this chapter after this initial classification. However, we have a detail analysis and proposed technique of same category discussed in the upcoming chapter.

i. Data Analysis Approaches (Zou & Xiao, 2008)

ii. Novel Data Filling Approach for an Incomplete Soft Set (DFIS) (H. Qin, et al., 2012a)

iii. An efficient decision making approach in incomplete soft set (Kong et al., 2014)

3.2.3 Analysis of the Pre-Processed Incomplete Soft sets

The techniques of pre-processed category for the capability of number of possible re-calculable values are analyzed. Their maximum possible re-calculating limits are checked and it is focused whether these techniques in their current form can be used for re-calculating overall missing values from available parity bits, supported sets and diagonal aggregates. For this purpose, each technique of this category is considered one by one and both capabilities are generalized after individual analysis.

3.2.3.1 Using Parity Bits and Supported Set

This technique uses two sets of parity bits each for columns and rows and one set of supported values for each row. If the size of soft set is $m \times n$ where *m* is the number of rows and *n* is the number of columns in the table, then there are $m+n$ number of parity bits and *m* number of supported values in the supported set. It is analyzed that, like simultaneous linear equations, the capability of one parity bit is re-calculation of only one unknown and hence $m+n$ parity bits can re-calculate $m+n$ unknowns only. In general, the capability of one supported value is also calculating one unknown; hence *m* number of supported values can calculate *m* unknowns only. To combine both capabilities, up to $2m+n$ unknowns can be calculated through this technique. This capability has also some boundaries that the $2m+n$ unknowns should in proper order otherwise the technique is unable to recalculate them. Without going to further details and focusing on overall missing data recalculation, this research gives a hint for the mentioned order that the technique is ideal if missing values are only two per row and one per column. In some special cases, supported set can calculate more number of unknowns than its values. The capability of this technique is explained in below example with its special case. It is notable that such special cases can't be generalized for extending the capability until predicted correctly and the general capability are considered as the exact capability of this approach.

Example 3.1: Consider Table 3.1 having the number of rows $m = 10$ and number of columns $n = 6$. The size of soft set is $m \times n = 60$ and the number of unknowns is 40 as shown by $*_1$ to $*_{40}$. As the capability of the technique is $2m + n = 26$ for this case, therefore are unable to recalculate 40 unknowns through it. As there is no single value missing in any row or column therefore parity bit can't be used here for re-calculating any single value. However, a special case of supported value is observed for object u_2 that $\text{supp}(2) = 5$ therefore, $*_1 = *_2 = *_3 = *_4 = *_5 = 1$

Table 3.1: Incomplete Soft Set of size 60 with 40 unknowns

U/E	e_1	e_2	e_3	e_4	e_5	e_6	P_{bit}	$Supp$
u_1	1	0	1	0	0	0	0	2
u_2	0	$*_1$	$*_2$	$*_3$	$*_4$	$*_5$	1	5
u_3	1	$*_6$	$*_7$	$*_8$	$*_9$	$*_{10}$	0	4
u_4	1	$*_{11}$	$*_{12}$	$*_{13}$	$*_{14}$	$*_{15}$	0	4
u_5	0	$*_{16}$	$*_{17}$	$*_{18}$	$*_{19}$	$*_{20}$	1	1
u_6	$*_{21}$	$*_{22}$	$*_{23}$	$*_{24}$	$*_{25}$	$*_{26}$	1	1
u_7	0	$*_{27}$	$*_{28}$	$*_{29}$	$*_{30}$	$*_{31}$	1	1
u_8	1	$*_{32}$	$*_{33}$	$*_{34}$	$*_{35}$	$*_{36}$	1	5
u_9	1	1	1	$*_{37}$	$*_{38}$	$*_{39}$	1	5
u_{10}	$*_{40}$	0	0	1	0	0	0	2
C_{bit}	0	1	0	1	1	1	--	

In addition to above, there are few special small cases for which this technique is always capable of recalculating overall missing values. The general case of condition is that when $2m+n \geq m \times n$. For example, for a 3×3 table $2m+n = m \times n = 9$ and for 2×3 table $2m+n = 7 > m \times n = 6$. Hence, in general, if the size of table is bigger than the capability of this approach, the approach can't be used for overall missing values.

3.2.3.2 Using rows, columns and diagonals aggregates

This technique is relatively more powerful compare to previous. The reason is very simple that it uses more simultaneous linear equations and makes its capability slightly increased. In addition to previous approach, this approach uses supported sets of both rows and columns and its capability for only rows and columns is $m+n$. It also uses left to right and right to left aggregates of diagonals and the number of diagonals in a table is $m+n-1$. For both directional diagonals, the re-calculating capability becomes double i.e. $2(m+n-1)$ and combining it with rows and columns capability, it becomes $3m+3n-2$. Hence, this technique is also not capable of overall missing values recalculation except few special cases when $3m+3n-2 \geq m \times n$.

3.2.3.3 Overall missing values recalculation

It is clear from above analysis that existing techniques of pre-processed incomplete soft set cannot be used for overall all missing values recalculation in their current form. Therefore, this study extends it and proposes another technique in the upcoming section that is able to do it.

3.3 Entire Missing Values Recalculation from Available sets of Aggregates

In this section, the concept of recalculating the entire BIS values from available aggregates is presented. First, the question, "Is finding more variables than available relations through linear equations possible?" is answered. After answering this for a special case of BIS, the proposed method with several important definitions and examples is presented.

3.3.1 Solving non-simultaneous linear equations in real domain

Simultaneous linear equations are defined as, "The set of two or more than two equations is called the set of simultaneous linear equations or simply simultaneous linear equations, if each equation contains two or more variables, such that the number of variables is less than or equal to the number of equations, and the values of variables can satisfy both or all equations simultaneously."

Suppose there is a set of linear equations as follows:

$$x + y + z = 2 \tag{3.1}$$

$$z = 1 \tag{3.2}$$

According to the above definition, this set is not the set of simultaneous linear equations because the number of relations is less than the number of variables and an exact solution of unknowns cannot be found. If $z = 1$ in Equation (3.1) the following relation is obtained:

$$x + y = 1 \tag{3.3}$$

Infinite number of values for x and y of relation (3.3) can be identified in the real domain. Thus, the sum of both will be equal to 1. In the case of real numbers, finding exact values through non-simultaneous linear equations is impossible.

3.3.2 Solving non-simultaneous linear equations in Boolean domain

The set of linear equations given above is reconsidered. If it is known that the domain of these variables is of Boolean values, then two possible solutions for the relations (3.3) can easily be identified as given below.

i. $x = 1$ and $y = 0$

ii. $x = 0$ and $y = 1$

Two steps are involved in finding the above solutions. First, suppose $x = 0$ and place it in (3.3) to obtain $y = 1$. Then, supposing $x = 1$ yields $y = 0$. Hence, unlike the previous case of real domain, obtaining the finite number of possible values for such non-simultaneous relations by supposition in the binary domain is possible. If there is a clue of cross confirmation to select either one of the possible result or the other, then the exact one solution among all possible solutions can be identified.

3.3.3 Possibility of finding entire missing values in Boolean-valued information system from aggregates

From the above discussion, the following points can be concluded as follows.

1. If there is a finite domain of values, obtaining all possible values of unknowns is possible even through the non-simultaneous linear equation by supposition.
2. If there is a clue of cross confirmation, then one exact set of values for unknowns among the set of all possible values calculated in Step 1 can be selected.

Accordingly, BIS has the following:

a. A finite domain of binary values, and either 0 or 1 can be supposed as the possible value to obtain all possible sets of values.
b. Four sets of aggregates, where one is selected as the linear equation for the supposition of Step 1, and the other three sets function as the clue of cross confirmation for selecting one set of values as the exact solution.

Hence, recalculating all missing values from the aggregates in BIS is possible.

3.3.4 Proposed Method

The main idea of the proposed method is concluded in the above points a and b. To formalize the concept, several important definitions and algorithm are presented, and then an example is solved using the proposed algorithm as a proof of concept. Each LR and RL diagonals have two cases but this study defines one general case for those cases as follows.

Definition 3.1: Let (F,E) be a soft set and the diagonal be defined as $Diag_l = f(u_i,\ a_j)$, where $l = 1,2,\ldots,D,\ldots,2D-1,2D$, such that $D = m+n-1$, $m = |U|$, and $n = |A|$ are number of rows and columns, respectively.

From Definition 3.1, the concept of empty, universal, and hybrid (EUH) diagonals is introduced.

Definition 3.2: Let (F,E) be a soft set. A diagonal is called empty if its aggregate is equal to zero, i.e.

$$\sum f(u_i,\ a_j) = 0\ .$$

Definition 3.3: Let (F,E) be a soft set. A diagonal is called universal if its aggregate is equal to the number of its cells, i.e.

$$\sum f(u_i,\ a_j) = |f(u_i,\ a_j)|\ .$$

Definition 3.4: Let (F,E) be a soft set. A diagonal is called hybrid if it is neither empty nor universal, i.e.

$$0 < \sum f(u_i,\ a_j) < |f(u_i,\ a_j)|\ .$$

In several special cases, only empty and universal diagonals are used to calculate missing data without going to any supposition from hybrid diagonals. This makes the proposed approach more efficient, and the proposed algorithm successfully ends on Step 6. In most cases of large tables, it is impossible to accomplish this task on the bases of empty and universal diagonals only. Thus, it is needed to suppose binary values for hybrid diagonals.

Let $\sum f(u_i, a_j) = H_l$ be the aggregate value and $\left|f(u_i, a_j)\right| = M_l, \forall\left|a_{ij}\right| = 1$ be the cardinality or maximum value or size of a hybrid diagonal $Diag_l$.

Definition 3.5: Let (F, E) be a soft set. If S_l is the number of suppositions for diagonal $Diag_l$, then

$$S_l = \prod M_l \text{ , } l = 1,2,\ldots,D,\ldots,2D\text{-}1,2D .$$

Definition 3.6: Let (F, E) be a soft set. The total number of 1s in S_l for a $Diag_l$ must be H_l while the number of 0s will be automatically $M_l\text{-}H_l$.

In proposed approach, $m \times n$ table is constructed from the given number of rows and columns. All empty and universal diagonals are filled up according to Definitions 3.2 and 3.3 with 0s and 1s, respectively. Then, all columns, rows, and diagonals are checked and filled in if possible according to its aggregate values. Second, data is temporarily filled in the shortest diagonals first by supposing diagonal cells as 0 or 1 according to Definition 3.5. Suppositions are cross-checked with related aggregate values, where possible. Initially supposed values are permanently assigned to specific cells only if other aggregates verify it. Otherwise, the supposition order is changed. The process is repeated again until the original values are identified. These values are assigned permanently after confirmation of having no contradiction with any of the related aggregate. Proposed algorithm is given in Figure 3.1 for recalculating the entire BIS from aggregate values.

```
Calculating entire BIS from the aggregate values

Input: Rows, columns, and diagonals aggregates.

Output: Entire values of BIS.

1.  Calculate the table size from the cardinality of rows and
    columns aggregate sets.
2.  Assign a separate variable to each cell.
3.  Calculate null diagonals and set 0 to all its cells.
4.  Calculate universal diagonals and set all its cells to 1.
5.  Set values to the missing cells of rows, columns, and
    diagonal, according to their aggregate until the entire table
    is filled.
6.  End if all data is filled and all aggregates are satisfied;
    otherwise, move to next step.
7.  Assign values from 0 and 1 to the shortest incomplete
    diagonals, such that the aggregate value is satisfied.
8.  Assign values to the missing cells of rows, columns, and
    diagonal, according to their aggregate until the complete
    table is filled or any aggregate is disproved.
9.  Set permanently assigned values if these are not
    contradicting to any aggregate.
10. End if all missing data are filled; otherwise, go to Step 11.
11. Change the order of supposition and go to Step 7.
```

Figure 3.1: Algorithm for entire Boolean values recalculation from aggregates

The following example describes how the proposed algorithm handles missing data.

Example 3.2:

Supposing that there are four non-empty sets as given below,

1. $R_i = \{5,3,4,3,3,4,5\}$ represents the row aggregate values of the soft set.
2. $C_j = \{4,3,3,5,3,5,4\}$ represents the column aggregate values of the soft set.
3. $LR_D = \{1,2,0,2,3,2,4,3,4,3,1,1,1\}$ represents the LR diagonal aggregate values of the soft set.

4. $RL_D = \{1,1,2,2,2,3,3,5,3,1,3,0,1\}$ represents the RL diagonal aggregate values of the soft set.

Our target is to calculate all the soft set BIS entries from this data through the proposed approach.

Solution: $|R_i| = |C_j| = 7$ means that there are seven objects and seven parameters. Let (F,E) be the required soft set. $U = \{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$ and $E = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7\}$ are the object and parameter sets, respectively. A table of 7×7 order is constructed in Table 3.2 with rows representing the objects of the universal Set *U* and the columns representing the parameter Set *E*. All values are initially represented by * because they are unknown.

Table 3.2: Representation of unknown (F,E)

U/E	e_1	e_2	e_3	e_4	e_5	e_6	e_7
o_1	*	*	*	*	*	*	*
o_2	*	*	*	*	*	*	*
o_3	*	*	*	*	*	*	*
o_4	*	*	*	*	*	*	*
o_5	*	*	*	*	*	*	*
o_6	*	*	*	*	*	*	*
o_7	*	*	*	*	*	*	*

Another table (Table 3.3) is constructed and all unknowns values are assigned to temporary variables for identification, such that $O_i = \{s_i, t_i, v_i, w_i, x_i, y_i, z_i\}$ for $i = 1, 2, \ldots, 7$. The row and column aggregates are also shown in the same table.

Table 3.3: Representation of unknowns by variables with row and column aggregates

U/E	e_1	e_2	e_3	e_4	e_5	e_6	e_7	R_i
o_1	s_1	s_2	s_3	s_4	s_5	s_6	s_7	$R_1 = 5$
o_2	t_1	t_2	t_3	t_4	t_5	t_6	t_7	$R_2 = 3$
o_3	v_1	v_2	v_3	v_4	v_5	v_6	v_7	$R_3 = 4$
o_4	w_1	w_2	w_3	w_4	w_5	w_6	w_7	$R_4 = 3$
o_5	x_1	x_2	x_3	x_4	x_5	x_6	x_7	$R_5 = 3$
o_6	y_1	y_2	y_3	y_4	y_5	y_6	y_7	$R_6 = 4$
o_7	z_1	z_2	z_3	z_4	z_5	z_6	z_7	$R_7 = 5$
C_j	$C_1 = 4$	$C_2 = 3$	$C_3 = 3$	$C_4 = 5$	$C_5 = 3$	$C_6 = 5$	$C_7 = 4$	--

In Tables 3.4 and 3.5, this unknown table is shown with LR and RL diagonal aggregates.

Table 3.4: LR diagonal aggregate representation of unknown (F,E)

	LR	e_1	e_2	e_3	e_4	e_5	e_6	e_7	U
		s_1	s_2	s_3	s_4	s_5	s_6	s_7	o_1
$\mathrm{Diag}_{\mathrm{LR}}(\mathrm{k})$	$LR_1 = 1$	t_1	t_2	t_3	t_4	t_5	t_6	t_7	o_2
	$LR_2 = 2$	v_1	v_2	v_3	v_4	v_5	v_6	v_7	o_3
	$LR_3 = 0$	w_1	w_2	w_3	w_4	w_5	w_6	w_7	o_4
	$LR_4 = 2$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	o_5
	$LR_5 = 3$	y_1	y_2	y_3	y_4	y_5	y_6	y_7	o_6
	$LR_6 = 2$	z_1	z_2	z_3	z_4	z_5	z_6	z_7	o_7
	$LR_7 = 4$		▨			▨	🟦		
$\mathrm{Diag}_{\mathrm{LR_u}}(\mathrm{k})$	$LR_8 = 3$	▨			▨	🟦			
	$LR_9 = 4$			▨	🟦				
	$LR_{10} = 3$		▨	🟦					
	$LR_{11} = 1$	▨	🟦						
	$LR_{12} = 1$	🟦							
	$LR_{13} = 1$								

Table 3.5: RL diagonal aggregate of unknown (F,E)

U/E	e_1	e_2	e_3	e_4	e_5	e_6	e_7	RL	
o_1	s_1	s_2	s_3	s_4	s_5	s_6	s_7		
o_2	t_1	t_2	t_3	t_4	t_5	t_6	t_7	$RL_1=1$	$Diag_{RL}(k)$
o_3	v_1	v_2	v_3	v_4	v_5	v_6	v_7	$RL_2=1$	
o_4	w_1	w_2	w_3	w_4	w_5	w_6	w_7	$RL_3=2$	
o_5	x_1	x_2	x_3	x_4	x_5	x_6	x_7	$RL_4=2$	
o_6	y_1	y_2	y_3	y_4	y_5	y_6	y_7	$RL_5=2$	
o_7	z_1	z_2	z_3	z_4	z_5	z_6	z_7	$RL_6=3$	
								$RL_7=3$	
								$RL_8=5$	$Diag_{RL_u}(k)$
								$RL_9=3$	
								$RL_{10}=1$	
								$RL_{11}=3$	
								$RL_{12}=0$	
								$RL_{13}=1$	

Tables 3.4 and 3.5 show that LR_1, LR_2, LR_{13}, RL_1, RL_{11}, and $RL_{13}$are universal while LR_3 and LR_{12} are null. According to Definitions 3.2 and 3.3, the cells of universal diagonals are replaced with 1 and those of empty diagonals are replaced with zero. Some missing information as provided in Table 3.6 obtained from EUH.

Table 3.6: Incomplete table after null and universal diagonal filling

U/E	e_1	e_2	e_3	e_4	e_5	e_6	e_7
o_1	1	1	0	s_4	s_5	s_6	1
o_2	1	0	t_3	t_4	t_5	t_6	t_7
o_3	0	v_2	v_3	v_4	v_5	v_6	v_7
o_4	w_1	w_2	w_3	w_4	w_5	w_6	w_7
o_5	1	x_2	x_3	x_4	x_5	x_6	x_7
o_6	0	1	y_3	y_4	y_5	y_6	y_7
o_7	1	0	1	z_4	z_5	z_6	1

In Table 3.6, 1st column can be completed by placing $w_1 = 0$, thus Table 3.7 is obtained because it is known that $C_1 = 4$.

Table 3.7: Incomplete soft set after filling 1st column

U / E	e_1	e_2	e_3	e_4	e_5	e_6	e_7
o_1	1	1	0	s_4	s_5	s_6	1
o_2	1	0	t_3	t_4	t_5	t_6	t_7
o_3	0	v_2	v_3	v_4	v_5	v_6	v_7
o_4	0	w_2	w_3	w_4	w_5	w_6	w_7
o_5	1	x_2	x_3	x_4	x_5	x_6	x_7
o_6	0	1	y_3	y_4	y_5	y_6	y_7
o_7	1	0	1	z_4	z_5	z_6	1

Considering Table 3.7, and starting the supposition from the shortest incomplete diagonals, which are LR_{12} and RL_2. Both have two cells and aggregate values that are equal to 1. In both diagonals, one value must be 0 and the other must be 1. Supposing $y_7 = 0 = t_7 \Rightarrow z_6 = s_6 = 1$, the process cannot be proceeded without further supposition for the next shortest diagonals, which are LR_{11} and RL_3. These diagonals have three cells and aggregate values that are equal to 1 and 2, respectively. Supposing $x_7 = y_6 = v_7 = 0 \Rightarrow z_5 = t_6 = s_5 = 1$, Table 3.8 is obtained after placing these values.

Table 3.8: Placing non-contradicting supposed values for LR_{12}, RL_2, LR_{11} and RL_3

U/E	e_1	e_2	e_3	e_4	e_5	e_6	e_7
o_1	1	1	0	s_4	1	1	1
o_2	1	0	t_3	t_4	t_5	1	0
o_3	0	v_2	v_3	v_4	v_5	v_6	0
o_4	0	w_2	w_3	w_4	w_5	w_6	w_7
o_5	1	x_2	x_3	x_4	x_5	x_6	0
o_6	0	1	y_3	y_4	y_5	0	0
o_7	1	0	1	z_4	1	1	1

C_7 disproves the supposition in Table 3.8. It cannot be obtained by placing $w_7 = 1$ only because it is known that its aggregate is equal to 4. Reconsidering Table 3.7, all suppositions are disproved through cross-checking except $y_7 = s_6 = z_5 = v_7 = t_6 = 1$, which implies that $z_6 = t_7 = x_7 = y_6 = s_5 = 0$, by supposing different possible combinations. Meanwhile, Table 3.9 is obtained from placing these values.

Table 3.9: Placing values of non-contradictive supposition

U/E	e_1	e_2	e_3	e_4	e_5	e_6	e_7
o_1	1	1	0	s_4	0	1	1
o_2	1	0	t_3	t_4	t_5	1	0
o_3	0	v_2	v_3	v_4	v_5	v_6	1
o_4	0	w_2	w_3	w_4	w_5	w_6	w_7
o_5	1	x_2	x_3	x_4	x_5	x_6	0
o_6	0	1	y_3	y_4	y_5	0	1
o_7	1	0	1	z_4	1	0	1

In Table 3.9, 1 can be easily placed 1 for s_4, z_4, v_6, w_6, x_6, and 0 for w_7 using R_1, R_7, C_6, and C_7, thereby obtaining Table 3.10.

Table 3.10: Placing values of s_4, z_4, v_6, w_6, x_6 and w_7

U/E	e_1	e_2	e_3	e_4	e_5	e_6	e_7
o_1	1	1	0	1	0	1	1
o_2	1	0	t_3	t_4	t_5	1	0
o_3	0	v_2	v_3	v_4	v_5	1	1
o_4	0	w_2	w_3	w_4	w_5	1	0
o_5	1	x_2	x_3	x_4	x_5	1	0
o_6	0	1	y_3	y_4	y_5	0	1
o_7	1	0	1	1	1	0	1

Substituting Table 3.10 into 3.11, $t_5 = 0$ and $y_5 = 1$ form $LR_{10} = 3$ and $LR_4 = 2$, respectively.

Table 3.11: Placing values of t_5 and y_5

U/E	e_1	e_2	e_3	e_4	e_5	e_6	e_7
o_1	1	1	0	1	0	1	1
o_2	1	0	t_3	t_4	1	1	0
o_3	0	v_2	v_3	v_4	v_5	1	1
o_4	0	w_2	w_3	w_4	w_5	1	0
o_5	1	x_2	x_3	x_4	x_5	1	0
o_6	0	1	y_3	y_4	0	0	1
o_7	1	0	1	1	1	0	1

In Table 3.11, from $RL_8 = 5$, implies that $v_2 = w_3 = x_4 = 1$, thereby obtaining Table 3.12.

Table 3.12: Placing values of v_2, w_3 and x_4

U/E	e_1	e_2	e_3	e_4	e_5	e_6	e_7
o_1	1	1	0	1	0	1	1
o_2	1	0	t_3	t_4	1	1	0
o_3	0	1	v_3	v_4	v_5	1	1
o_4	0	w_2	1	w_4	w_5	1	0
o_5	1	x_2	x_3	1	x_5	1	0
o_6	0	1	y_3	y_4	0	0	1
o_7	1	0	1	1	1	0	1

Considering Table 3.12, given that $LR_4 = 2$, hence, $t_3 = 0 \Rightarrow t_4 = 1$ because $R_2 = 3$. Also considering $C_2 = 3$, which implies that $w_2 = x_2 = 0$, thereby obtaining Table 3.13.

Table 3.13: Placing values of t_3, t_4, w_2 and x_2

U/E	e_1	e_2	e_3	e_4	e_5	e_6	e_7
o_1	1	1	0	1	0	1	1
o_2	1	0	0	1	1	1	0
o_3	0	1	v_3	v_4	v_5	1	1
o_4	0	0	1	w_4	w_5	1	0
o_5	1	0	x_3	1	x_5	1	0
o_6	0	1	y_3	y_4	0	0	1
o_7	1	0	1	1	1	0	1

In Table 3.13, given that $LR_5 = 3$, $RL_5 = 2$, and $RL_{10} = 1$, then $v_3 = 1$, $v_5 = 0$, and $y_3 = 0$, respectively, in Table 3.14.

Table 3.14: Placing v_3, v_5 and y_3

U/E	e_1	e_2	e_3	e_4	e_5	e_6	e_7
o_1	1	1	0	1	0	1	1
o_2	1	0	0	1	1	1	0
o_3	0	1	1	v_4	0	1	1
o_4	0	0	1	w_4	w_5	1	0
o_5	1	0	x_3	1	x_5	1	0
o_6	0	1	0	y_4	0	0	1
o_7	1	0	1	1	1	0	1

In Table 3.14, $v_4 = x_3 = x_5 = 0$ from R_3 and R_5 $w_5 = 1$ from LR_8. Calculating the remaining values for w_4 and y_4, thereby obtaining a complete Table 3.15.

Table 3.15: Complete table after missing values recalculation

U/E	e_1	e_2	e_3	e_4	e_5	e_6	e_7
o_1	1	1	0	1	0	1	1
o_2	1	0	0	1	0	1	0
o_3	0	1	1	0	0	1	1
o_4	0	0	1	0	1	1	0
o_5	1	0	0	1	0	1	0
o_6	0	1	0	1	1	0	1
o_7	1	0	1	1	1	0	1

Therefore, all unknowns are successfully calculated through the proposed approach in Table 3.15. Supposing that P_i are the parameters functions for $i = 1, 2, \ldots, 7$, then

$$(F, E) = \left\{ \begin{array}{l} P_1 = \{o_1, o_2, o_5, o_7\} \\ P_2 = \{o_1, o_3, o_6\} \\ P_3 = \{o_3, o_4, o_7\} \\ P_4 = \{o_1, o_2, o_5, o_6, o_7\} \\ P_5 = \{o_4, o_6, o_7\} \\ P_6 = \{o_1, o_2, o_3, o_4, o_5\} \\ P_7 = \{o_1, o_3, o_6, o_7\} \end{array} \right\}$$

as the required soft set.

3.4 Conclusion

In this chapter, the existing approaches to data prediction and re-calculation in incomplete soft sets as BIS are discussed. The previous approaches are categorized to PP and UP categories and it is shown that only preprocessed incomplete soft sets can be used for recalculation, and missing values can only be predicted in the UP category. A new concept for the recalculation of the entire BIS missing values from aggregates in the PP category is also presented. Proposed approach recalculates all missing values from the aggregates of available rows, columns, and diagonals by supposition and cross confirmation. The algorithm of proposed technique is presented and explained it with an example as a proof of concept. In the future, this new idea can be used in many applications of binary data in mathematics, computer science, and in the field of data compression at the binary level.

CHAPTER 4: DATA FILLING IN UNPROCESSED INCOMPLETE SOFT SET THROUGH STRONGEST ASSOCIATION BETWEEN PARAMETERS[5]

4.1 Introduction

Soft set theory proposed by Molodtsov is considered as a mathematical model for dealing with vague and uncertain data (Molodtsov, 1999). This theory is a standard as compare to existing theories such as fuzzy set, rough set, vague set and statistical approach for dealing with vague data because of its adequate of parameterization. Research in the soft set theory both theoretical and practical has been attracted many attentions, especially in the field of decision making. The first attempt in soft set decision making is introduced by Maji et al. (P. Maji et al., 2002). They presented soft set first application in decision making by representing it in Boolean table and defined its reduct set. Their work of reduct was improved by Chen et al., further improved by Kong et al. and sequentially by Ma et al. for decision making of sub-optimal choices and simplified approaches, respectively (Degang Chen et al., 2005; Kong et al., 2008; Ma et al., 2011). In parallel to these developments, researchers used soft set for handling daily life's uncertain data issues and applied it in verity of useful applications (Cagman & Enginoglu, 2012; Naim Cagman, Serdar Enginoglu, & Filiz Citak, 2011; Çelik & Yamak, 2013; Herawan & Deris, 2011; Jun et al., 2009; Jun & Park, 2008; Kalaichelvi & Malini, 2011b; Kalayathankal & Singh, 2010; Sutoyo et al., 2016; Tanay & Kandemir, 2011; Xiao et al., 2009; Yuksel et al., 2013). But in some applications, researchers faced problem of incomplete soft set cases with partially missing values. Soft and its related sets data can be missed due to many factors such as improper entry, viral attack, security reasons and errors during data transfer. Incomplete soft sets can be no longer applied in any application or may yield extra-large, very small, unexpected

[5] The main idea of this chapter has already been published in ISI indexed journal "SpringerPlus" with the title "An alternative data filling approach for prediction of missing data in soft sets (ADFIS)"

and misleading results, if still applied. Such results, especially a wrong decision making can cause a huge loss to an individual or organizations. For coping with this situation, Zou et al. presented their techniques of weighted-average for calculating decision values and average probability for prediction of missing values in soft set and fuzzy soft set respectively (Zou & Xiao, 2008). Qin et al. proposed DFIS where it indicated that data prediction in incomplete soft set is more reliable and accurate if recalculated through association between parameters and they used simple probability for cases having zero or weak association (H. Qin, X. Ma, T. Herawan, & J. M. Zain, 2012b). Rose et al. also contributed in completion of incomplete soft set using parity bits and aggregate values (Mohd Rose et al., 2011; Rose et al., 2011). Sub-sequentially, Kong et al. (Kong et al., 2014) improved Zou et al. (Zou & Xiao, 2008) approach of incomplete soft set by presenting an equivalent probability technique having less complexity and also determining actual missing data instead of only decision values determination.

In previous chapter, the above mentioned techniques of handling incomplete soft set were classified into PP and UP categories. PP category techniques were discussed in detail and a new technique of entire missing values recalculation from aggregates was presented in the previous chapter. This chapter discusses the techniques of UP category. Existing techniques in UP category (Kong et al., 2014; H. W. Qin et al., 2012; Zou & Xiao, 2008) are explained one by one in Literature review chapter.

In this chapter, all exiting approaches of UP category are compared in term of accuracy, computational complexity and data integrity, and DFIS is found as most suitable among them for predicting missing values. An alternative data filling approach for prediction of missing data in soft sets is proposed. In summary, the contribution of this chapter is described as follow:

a. DFIS is indicated as most suitable for data prediction in UP incomplete soft set.

b. An alternative data filling approach is proposed that predicts incomplete data in UP soft set through strongest association unlike DFIS.

c. Extensive experiment tests on 04 UCI benchmark and causality workbench lung cancer (LUCAP2) data sets are performed to validate proposed work and to show the performance of proposed approach. Both, proposed approach and DFIS are implemented in MATLAB and the results are compared.

4.2 Analysis of previous approaches in UP category

In this section, previous approaches of UP category in incomplete soft set are discussed. They are analyzed for indicating most suitable technique for finding missing data in incomplete soft set.

4.2.1 Previous approaches of UP category

Previous approaches of this category are discussed in chapter 2 in details and an incomplete soft set of Table 4.1, Example 2.4 is completed using each technique. However, only results and key points of each technique are mentioned here for analysis and avoiding repetition.

4.2.1.1 Zou et al. approach

Incomplete Example 2.4 completed through Zou et al. approach (Zou & Xiao, 2008) is given in Table 4.1 and their main points are given below.

a) Uses weighted average technique for finding decision values.

b) Finds decision values only while actual missed values still remain missed.

c) Computational complexity is $O(n.2^n)$ (Kong et al., 2014).

d) Accuracy of decision values is low (H. Qin, X. Ma, et al., 2012b).

Table 4.1: Incomplete soft set Example 2.4 completed through Zou et al. approach

U/E	e_1	e_2	e_3	e_4	e_5	e_6	d_i
s_1	0	1	1	1	0	0	3
s_2	0	1	0	0	0	1	2
s_3	1	0	0	1	0	0	2
s_4	1	0	$*_1$	0	$*_2$	1	2.57
s_5	0	1	1	0	0	1	3
s_6	1	0	0	$*_3$	0	0	1.43
s_7	$*_4$	1	1	1	0	0	3.43
s_8	0	0	1	0	0	1	2

4.2.1.2 DFIS

Incomplete soft set Example 2.4 completed through DFIS (H. Qin, X. Ma, et al., 2012b) is given in Table 4.2 and main points of this technique are given below.

a) Uses association between parameters for data filling a give second priority to probability in case of weak association

b) Accuracy of decision values is high compare to Zou et al. approach.

c) Assigns values to actual missing values as well unlike Zou et al. approach

d) Easy to understand and implement as compare to Zou et al. approach

Table 4.2: Incomplete Example 2.4 completed using DFIS

U/E	e_1	e_2	e_3	e_4	e_5	e_6
s_1	0	1	1	1	0	0
s_2	0	1	0	0	0	1
s_3	1	0	0	1	0	0
s_4	1	0	1	0	0	1
s_5	0	1	1	0	0	1
s_6	1	0	0	1	0	0
s_7	0	1	1	1	0	0
s_8	0	0	1	0	0	1

4.2.1.3 Kong et al. approach

Incomplete soft set of Example 2.4 completed through Kong et al. approach (Kong et al., 2014) is given in Table 4.3 and their main points are given below.

a) Uses probability for finding decision values.

b) Assigns rational values to actual missed values as well which affects the integrity of standard soft set.

c) Computational complexity is $O(n^2)$.

d) Very easy to understand and implement.

Table 4.3: Incomplete soft set of Example 2.4 completed using Kong et al. approach

U/E	e_1	e_2	e_3	e_4	e_5	e_6	d_i
s_1	0	1	1	1	0	0	3
s_2	0	1	0	0	0	1	2
s_3	1	0	0	1	0	0	2
s_4	1	0	$\frac{4}{4+3}$	0	$\frac{0}{0+7}$	1	2.57
s_5	0	1	1	0	0	1	3
s_6	1	0	0	$\frac{3}{3+4}$	0	0	1.43
s_7	$\frac{3}{3+4}$	1	1	1	0	0	3.43
s_8	0	0	1	0	0	1	2

4.2.2 Indication of most suitable approach among existing techniques in UP category

As Zou et al. and Kong et al. approaches have same results (Kong et al., 2014) and Zou et al. approach is compared with DFIS with details (H. Qin, X. Ma, et al., 2012b). To conclude, below associative way is adopted for comparing all three previous techniques.

4.2.2.1 Zou et al. approach versus Kong et al. approach

As Zou et al. approach calculates only decision value of incomplete soft set and the missing data remains still missing. While, Kong et al. approach has same results of d_i as that of Zou et al. approach along with assigning a set of values to originally missed information. Secondly, the computational complexity of Kong et al. approach is $O(n^2)$ while that of Zou et al. approach is $O(n.2^n)$ showing that Kong et al. approach is less complex compare to Zou et al. approach (Kong et al., 2014). Therefore, Kong et al. technique is more appropriate and efficient than Zou et al. approach.

4.2.2.2 Kong et al. approach versus DFIS

As Kong et al. approach works only on probability, ignoring any association between parameters might result probably in different values from actual. Secondly, it predicts missing values in [0, 1] range, while the actual value must be either 0 or 1 in standard soft set (Boolean information system). In contrast, DFIS prefer to predict actual values through association and use probability when the association is not strong. Secondly, in both cases, it calculates binary values maintaining the integrity of standard soft set. Thirdly, compare to Zou et al. results; its decision values results are much closer to actual values as shown in experimental results (H. Qin, X. Ma, et al., 2012b). The average of Mean Absolute Percentage Error (MAPE) of DFIS is 0.07, while that of Zou et al. approach is 0.11 for all five data sets used in DFIS. If this average of MAPE is converted to percent accuracy of both approaches, then the average accuracy of DFIS is 93.17% while that of Zou et al. approach is 89.12% in calculating decision values. It is notable that Zou et al. and Kong et al. approaches have same results of decision values (Kong et al., 2014); consequently, the average accuracy of DFIS in decision values comes to be 4.04% higher than Kong et al. technique.

Moreover, the computational complexity of DFIS is calculated which consists of below steps.

1. Access whole data set of $m \times n$ size once for getting the number of missing values
2. Compute the degrees of consistencies and inconsistencies of complexity n
3. Compute probability of n complexity when the association is weak
4. Access once again $m \times n$ table for inserting the computed values

Combining all, results in $m \times n + n + n + m \times n = 2.m.n + 2.n$. Supposing $m = n$ and considering big O notation, then $2.m.n + 2.n = 2..n^2 + 2.n \geq 2..n^2 \geq n^2$ for larger values of n. Hence, the complexity of DFIS is $O(n^2)$, which is equal to the complexity of Kong et al. approach. Hence DFIS is more suitable than Kong et al. approach.

4.2.2.3 DFIS as the most suitable technique among existing UP incomplete soft set

In above associative comparison, it is shown that Kong et al. technique is better than Zou et al. technique and DFIS is better than Kong et al. technique. Therefore, DFIS is most appropriate for missing data prediction in soft set among all three previous approaches. This comparison is summarized in Table 4.4 as follow:

Table 4.4: Comparison of Unprocessed incomplete soft set handling approaches

Advantages|Techniques	Zou et al. approach	Kong et al. approach	DFIS
Calculates missing value	No	Yes	Yes
Less complexity	No	Yes	Yes
Use association between parameters	No	No	Yes
Calculates Binary values (Standard soft set)	No	No	Yes
Accuracy is high	No	No	Yes

4.2.3 Problems of DFIS

Above comparison illustrates that DFIS is most suitable for prediction of missing values in unprocessed incomplete soft set. It is because DFIS prefers association between parameters for the prediction and give second priority to probability. But the accuracy of DFIS is not 100% as shown in the results section of the respective article.

The MAPE and percent accuracy already discussed in associative comparison is for decision values. The MAPE of DFIS is 0.07 while the derived accuracy from MAPE is 93.17%. Although accurate decision has direct relation with accuracy of actual data yet the actual accuracy of predicted data cannot be found in the literature. The reason for unavailability of DFIS actual accuracy is that the baseline technique for DFIS is Zou et al. technique and Zou et al. do not have actual data for comparison rather they have the decision values only. DFIS has no option to compare actual results with Zou et al. approach and they were bound to calculate the decision values from predicted actual values. The low accuracy of DFIS is obviously understandable from this situation but it is felt that the average range of their accuracy for actual predicted values (other than decision values) needs to be explained for more visibility. Therefore, DFIS is implemented in MATLAB and values are predicted through it after deletion from certain benchmark data sets. Accuracy of DFIS is given in Table 4.5 while further details and measures of these data sets and experiments are explained later under results section in the proposed approach of this chapter.

Table 4.5: Average accuracy of DFIS for benchmark data sets calculated after deletion of values and recalculating through DFIS in MATLAB

Data Sets	Percent accuracy of DFIS
Zoo Data Set	81.26
Flags Data Set	74.02
SPECT Hearts Data Set	76.41
Congressional Votes Data Set	65.50
LUCAP2 Data Set	71.61
Average	73.76

Table 4.5 shows that DFIS itself has low accuracy problem and there might exist the chances of its accuracy improvement. This study has observed that the algorithm used by DFIS does not consider strongest association between parameter and if this reliable association is included in DFIS the results will be more accurate. Therefore, DFIS is modified for better prediction accuracy and an alternate data filling approach is presented which predict missing values in incomplete soft set through considering strongest association between parameters.

4.3 Proposed Approach

In this section an alternative approach for data filling of incomplete soft sets is presented. The technique is explained with the help of definitions, mathematical relations, algorithm and step by step procedure using a practical example as a proof of concept followed by experimental results and discussion

4.3.1 Materials and methods of proposed technique

. The previous approach DFIS preferred association between parameters to predict missing values than probability and this study has discussed that association results in more accurate values than probability. But DFIS itself is unable to precisely consider all possible associations for getting more accurate results. In contrast to DFIS, proposed approach revises the association calculating method to consider all possible associations precisely and predict maximum possible number of unknowns through it. The novelty of proposed approach is that, it relies on strongest association unlike DFIS.

DFIS uses below mathematical relations for finding consistent and inconsistent associations and the degree of consistency and inconsistency as explained in the literature review of DFIS.

$$CN_{ij} = \left| \left\{ x \; \middle| F_{e_i}(x) = F_{e_j}(x),\; x \in U_{ij} \right\} \right| \tag{4.1}$$

$$CD_{ij} = \frac{CN_{ij}}{\left|U_{ij}\right|} \tag{4.2}$$

$$IN_{ij} = \left| \left\{ x \; \middle| F_{e_i}(x) \neq F_{e_j}(x),\; x \in U_{ij} \right\} \right| \tag{4.3}$$

$$ID_{ij} = \frac{IN_{ij}}{\left|U_{ij}\right|} \tag{4.4}$$

Above relations of DFIS are also used in proposed approach to find consistency (CN), consistency degree (CD), inconsistency (IN) and inconsistency degree (ID) between parameter i and j.

Below relation is defined, to find strongest association between all parameters.

$$SA_{ij} = \left| \max \left\{ \max \left\{ CD_{ij}, ID_{ij} \right\} \right\} \right| \tag{4.5}$$

where CD_{ij}, ID_{ij} are the degrees of consistencies and inconsistencies of each parameter *i* containing missing values with all other parameters *j* and SA_{ij} is the strongest association among all parameters, between parameter *i* (containing unknown) and (corresponding) parameter *j*. The following definition presents the notion of consistency between two parameters.

Definition 4.1: *Two parameters* e_i *and* e_j *are said to be consistent* $e_i \Leftrightarrow e_j$ *with each other if there is strongest association between them. i.e.* $SA_{ij} \geq \lambda$ *and* $\max\{CD_{ij}, ID_{ij}\} = CD_{ij}$, *where* λ *is a pre-set threshold values (for more details, see discussions)*

From Definition 4.1, it can be seen that if two parameters are consistent to each other, then its corresponding elements are also consistent with each other. If $e_i \Leftrightarrow e_j$ then $F(e)_{ni} \Leftrightarrow F(e)_{nj}$, if $F(e)_{ni} = *$ then

$$F(e)_{ni} = F(e)_{nj} \tag{4.6}$$

where, * is unknown and *n* is the object position (row) of parameter value $F(e)$. The following definition presents the notion of inconsistency between two parameters.

Definition 4.2: *Two parameters* e_i *and* e_j *are said to be inconsistent* $e_i \Rightarrow e_j$ *with each other if there is strongest inconsistent association between them. i.e.* $SA_{ij} \geq \lambda$ *and* $\max\{CD_{ij}, ID_{ij}\} = ID_{ij}$.

From Definition 4.2, it can be seen that if two parameters are inconsistent to each other, then its corresponding elements are also inconsistent with each other. If $e_i \Rightarrow e_j$ then $F(e)_{ni} \Rightarrow F(e)_{nj}$, if $F(e)_{ni} = *$ then

$$F(e)_{ni} = 1 - F(e)_{nj} \tag{4.7}$$

where, * is unknown and *n* is the object position (row) of parameter value $F(e)$. The following definition presents the notion of non-association between two parameters.

Definition 4.3: *Two parameters e_i and e_j are said to be non-associated $e_i \not\Leftrightarrow e_j$ if there exist no strongest association between them i.e. $SA_{ij} < \lambda$.*

From Definitions 4.1-4.3, proposed algorithm is derived as described in Figure 4.1.

```
Proposed Algorithm for data filling of incomplete soft set in
                          UP category

Input: Incomplete Soft Set

Output: Complete Soft Set

1  Find the columns i having unknown values ($F(e)_{ij} = *$).
2  Calculate strongest association ($SA_{ij}$).
3  Indicate k-th column having strongest association ($SA_{kj}$)
   with j-th column.
4  Select unknown(s) of k-th column only (Set $F(e)_{kj} = F(e)_{ij}$).
5  If $e_k \Leftrightarrow e_j$, put $F(e)_{nk} = F(e)_{nj}$.
6  If $e_k \Rightarrow e_j$, put $F(e)_{nk} = 1 - F(e)_{nj}$.
7  If $e_k \not\Leftrightarrow e_j$, calculate $n_1$ and $n_0$ for k-th column.
8  If $n_1 \geq n_0$, put $F(e)_{ik} = 1$.
9  If $n_1 < n_0$, put $F(e)_{ik} = 0$.
10 End if all missing values are predicted else go to step 1.
```

Figure 4.1: Proposed Algorithm for data filling of incomplete soft set in UP category

From above algorithm, the proposed approach firstly calculates the unknown(s) of the column having greatest association than all other columns among whole table. Before proceeding to further prediction, it inserts the recently calculated value(s) having strongest association in incomplete table. In next step, it again calculates association among parameters of whole table with consideration of the weight of recently inserted (most reliable) value(s) and finds strongest association again. The process of finding strongest association and predicting unknowns is repeated until all unknown data is filled or the condition of threshold disqualifies. In case of weak association, proposed approach uses simple comparison of n_1 and n_0 instead of calculating p_1 and p_0.

The main difference between DFIS and proposed method is that, DFIS calculates association among all parameters only once and decides on its base but proposed approach calculates it again and again after inserting the unknown value in one column being calculated through strongest association.

Proposed approach is further explained for understanding and comparison with DFIS in Example 4.1 with same incomplete case of Example 2.4.

Example 4.1: Prediction of unknowns for incomplete soft set case Example 2.1 through proposed approach. Consider Example 2.4 given in Table 4.6, for same case and same threshold value (λ=0.85).

Table 4.6: Incomplete soft set of Example 4.2

U/E	e_1	e_2	e_3	e_4	e_5	e_6
s_1	0	1	1	1	0	0
s_2	0	1	0	0	0	1
s_3	1	0	0	1	0	0
s_4	1	0	$*_1$	0	$*_2$	1
s_5	0	1	1	0	0	1
s_6	1	0	0	$*_3$	0	0
s_7	$*_4$	1	1	1	0	0
s_8	0	0	1	0	0	1

Step 1: Table 4.7 is constructed which contains the values of $\max\{CD_{ij}, ID_{ij}\}$.

Table 4.7: $\max\{CD_{ij}, ID_{ij}\}$ **:---(1)**

E^*/E	e_1	e_2	e_3	e_4	e_5	e_6
e_1	--	-0.86	-0.83	0.67	0.67	-0.83
e_3	-0.83	0.71	--	±0.5	-0.67	0.57
e_4	0.67	0.57	±0.5	--	±0.5	-1
e_5	0.67	-0.57	0.57	±0.5	--	0.57

From Table 4.7, according to equation (4.5) SA_{46}=1, for parameter 4 with parameter 6. As SA_{ij} >λ and $\max\{CD_{ij}, ID_{ij}\} = ID_{ij}$, definition 4.2 satisfies, therefore, $e_4 \Rightarrow e_6$

and $F(e)_{64} \Rightarrow F(e)_{66}$. In Table 4.6, $F(e)_{64} = *_3$ hence, $F(e)_{64} = 1 - F(e)_{66}$ according to equation (4.7). As $F(e)_{66} = 0$ in Table 4.6, implies that $F(e)_{64} = 1 - 0 = 1$. Hence, $*_3 = 1$. After putting this value, Table 4.8 is got as an updated case of incomplete data.

Table 4.8: Incomplete case after Inserting First Calculated Unknown ($*_3$) of Strongest Association

U/E	e_1	e_2	e_3	e_4	e_5	e_6
s_1	0	1	1	1	0	0
s_2	0	1	0	0	0	1
s_3	1	0	0	1	0	0
s_4	1	0	$*_1$	0	$*_2$	1
s_5	0	1	1	0	0	1
s_6	1	0	0	**1**	0	0
s_7	$*_4$	1	1	1	0	0
s_8	0	0	1	0	0	1

Step 2: Including the weight of recently calculated $*_3$ in Table 4.8, Table 4.9 is calculated containing the new values of $\max\{CD_{ij}, ID_{ij}\}$.

Table 4.9: $\max\{CD_{ij}, ID_{ij}\}$**: --- 2 for Updated Table 4.8**

D_{ij}	e_1	e_2	e_3	e_4	e_5	e_6
e_1	--	-0.86	-0.83	0.71	0.57	-0.71
e_3	-0.83	0.71	--	-0.57	-0.57	0.57
e_5	0.57	-0.57	-0.57	-0.57	--	0.57

In Table 4.9, the strongest association is that of e_1 with e_2, SA_{12}=|-0.86|>λ, similar to step 1, $*_3 = 0$ and updated Table 4.10 is obtained.

Table 4.10: Incomplete case after putting values of 1st and 2nd unknowns $*_3$ and $*_4$

U/E	e_1	e_2	e_3	e_4	e_5	e_6
s_1	0	1	1	1	0	0
s_2	0	1	0	0	0	1
s_3	1	0	0	1	0	0
s_4	1	0	$*_1$	0	$*_2$	1
s_5	0	1	1	0	0	1
s_6	1	0	0	**1**	0	0
s_7	**0**	1	1	1	0	0
s_8	0	0	1	0	0	1

Step 3: Based on updated Table 4.10, $\max\{CD_{ij}, ID_{ij}\}$ is calculated in Table 4.11 as follow.

Table 4.11: Calculation of $\max\{CD_{ij}, ID_{ij}\}$:--- 3 for updated Table 4.10

E^* / E	e_1	e_2	e_3	e_4	e_5	e_6
e_3	-0.86	0.71	--	-0.57	-0.57	0.57
e_5	0.71	-0.57	-0.57	-0.57	--	0.57

It can be observed from Table 4.11 that unlike DFIS, SA_{31}= |-0.86|>λ also entered into defined threshold range of association and first unknown $*_1 = 0$ getting updated incomplete case in Table 4.12.

Table 4.12: After putting values of $*_1, *_3$ and $*_4$

U/E	e_1	e_2	e_3	e_4	e_5	e_6
s_1	0	1	1	1	0	0
s_2	0	1	0	0	0	1
s_3	1	0	0	1	0	0
s_4	1	0	**0**	0	$*_2$	1
s_5	0	1	1	0	0	1
s_6	1	0	0	**1**	0	0
s_7	**0**	1	1	1	0	0
s_8	0	0	1	0	0	1

Step 4: The value of $\max\{CD_{ij}, ID_{ij}\}$ for Table 4.12 is recalculated in Table 4.13 as follow:

Table 4.13: Calculation of $\max\{CD_{ij}, ID_{ij}\}$:--- 4 for updated Incomplete Table 4.12

E^* / E	e_1	e_2	e_3	e_4	e_5	e_6
e_5	0.71	-0.57	-0.57	-0.57	--	0.57

As $SA_{51} = 0.71$ in Table 4.13 means $e_5 \not\Leftrightarrow e_1$ therefore, $*_2$ cannot be calculated through association for λ=0.85. This case is falling under Definition 4.3 and proposed approach uses probability for it. It can be seen from Table 4.12, that for e_5, $n_1 = 0$ and $n_0 = 7$. As $n_0 > n_1$ therefore, $*_2 = 0$. Hence, using proposed approach, all missing values are obtained in complete Table 4.14.

Table 4.14: Completed Soft Set using proposed method

U / E	e_1	e_2	e_3	e_4	e_5	e_6
s_1	0	1	1	1	0	0
s_2	0	1	0	0	0	1
s_3	1	0	0	1	0	0
s_4	1	0	**0**	0	**0**	1
s_5	0	1	1	0	0	1
s_6	1	0	0	**1**	0	0
s_7	**0**	1	1	1	0	0
s_8	0	0	1	0	0	1

4.3.2 Results

In this section, the improvement in accuracy of the predicted values in incomplete soft set using proposed approach is discussed. Firstly, the incomplete case in Example 2.4 is discussed for prediction results by DFIS and proposed method. Then, the results obtained from DFIS and proposed method for four UCI benchmark datasets and Causality workbench LUCAP2 data set are discussed. Some important discussions are provided after the results presentations and shortcomings of proposed method are also discussed in the end of this section.

4.3.2.1 Results from given example

Refer to comparison Table 4.15 of predicted unknowns, obtained from Table 4.2 and Table 4.14 using DFIS and proposed approach respectively. All values predicted through DFIS and proposed method are same except $*_1$, although the threshold is same for both approaches. $*_1$ got neither only complemented value for both techniques but also calculated through different ways i.e. through association in proposed approach and through probability by DFIS. The DFIS has proved that association is more reliable than probability; therefore, this study claims that the value of $*_1$ calculated as 0 using association by proposed approach is more accurate than predicted as 1 by DFIS using probability.

Table 4.15: Comparison of DFIS and proposed method predicted values for incomplete case of Example 2.4

	Predicted results through			
	DFIS		PROPOSED APPROACH	
Unknown	Value	Using	Value	Using
$*_1$	**1**	**Probability**	**0**	**Association**
$*_2$	0	Probability	0	Probability
$*_3$	1	Association	1	Association
$*_4$	0	Association	0	Association

Suppose an unknown predicted through association has 90% accuracy and that predicted through probability has 60%. Then the average accuracy of DFIS is 75% while that of proposed technique is 83% for this case as shown in the graph of Figure 4.2.

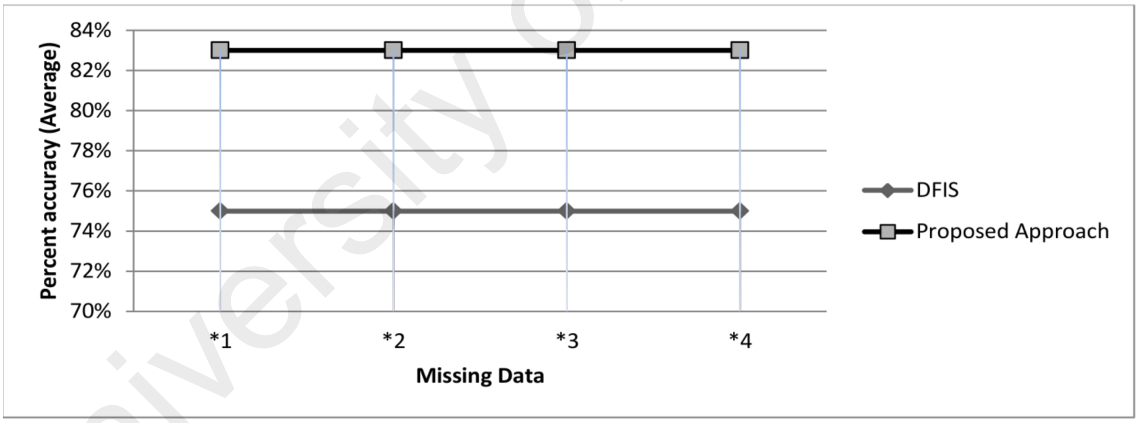


Figure 4.2: Performance comparison of DFIS and proposed approach for incomplete case of Example 2.4, Table 4.2

4.3.2.2 UCI Benchmark Data sets

Similar to DFIS (H. Qin, X. Ma, et al., 2012b), DFIS and proposed algorithm is tested, for four data sets from UCI benchmark database.

30 to 600 entries are randomly deleted ten times from Zoo, Flags, Congressional votes and SPECT hearts data sets and re-calculated it using both approaches by

implementing both algorithms in MATLAB. It is found that average accuracy of DFIS is 74.30% while that of proposed approach is 78.49% i.e. proposed algorithm performs 4.19% better than DFIS. Average performance graph is shown Figure 4.3. Further details and experimental results of each data set are individually discussed below.

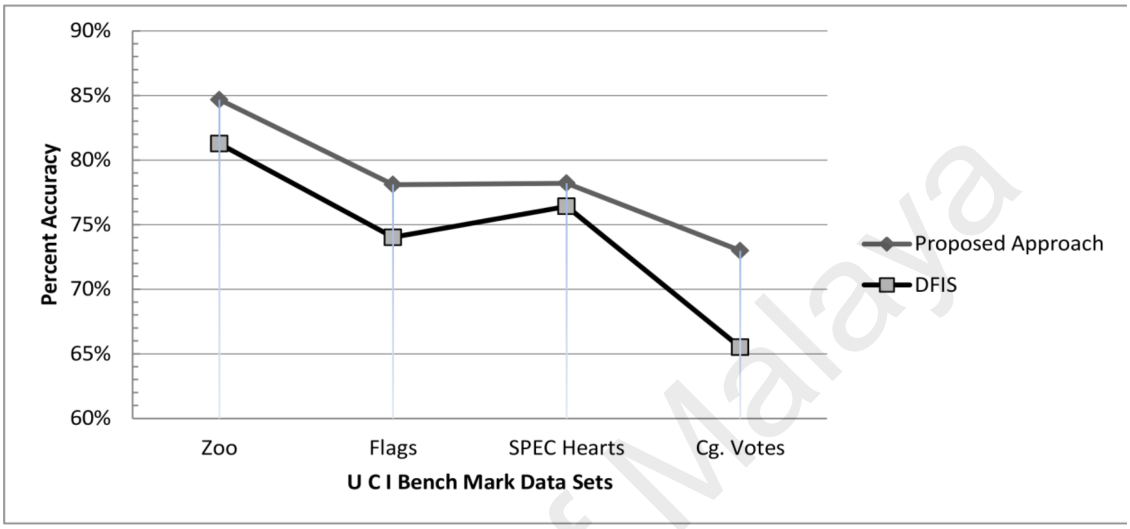


Figure 4.3: Average accuracy performance comparison of proposed method and DFIS for UCI Benchmark data sets

(a) ***Zoo data set***

Zoo data set contains 101 types of different animals with their 18 different features like presence of feather, teeth, backbone and hair. This study selected only 15 parameters having Boolean values and randomly deleted ten times the number of values 91, 87, 107, 91, 97, 98, 79, 82, 93 and 88 from it. All deleted values are recalculated using both approaches (DFIS and proposed). Percent accuracy graph of these results is given in Figure 4.4.

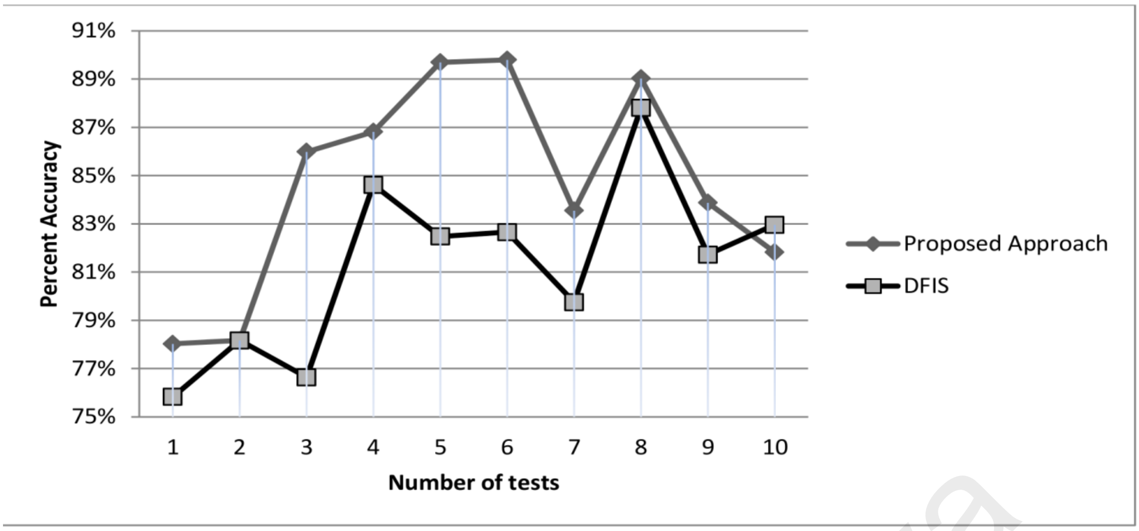


Figure 4.4: Percentage prediction accuracy for Zoo Data Set

Average performance of DFIS's accuracy is 81.26% while that of proposed method is 84.67% i.e. proposed method performs 3.41% accurate than DFIS for Zoo data set.

(b) ***Flags Data Set***

Flags dataset contains national flags description of 128 countries with 28 parameters. Out of all only 13 parameters are Boolean which are selected for the testing purpose. Accuracy graph for randomly deleted number of values 110, 43, 151, 92, 84, 151, 200, 538, 189 and 49 is given in Figure 4.5 for flag data set. Performance of proposed approach is 4.08% better than DFIS as DFIS average accuracy is 74.02% while that of proposed approach is 78.10%.

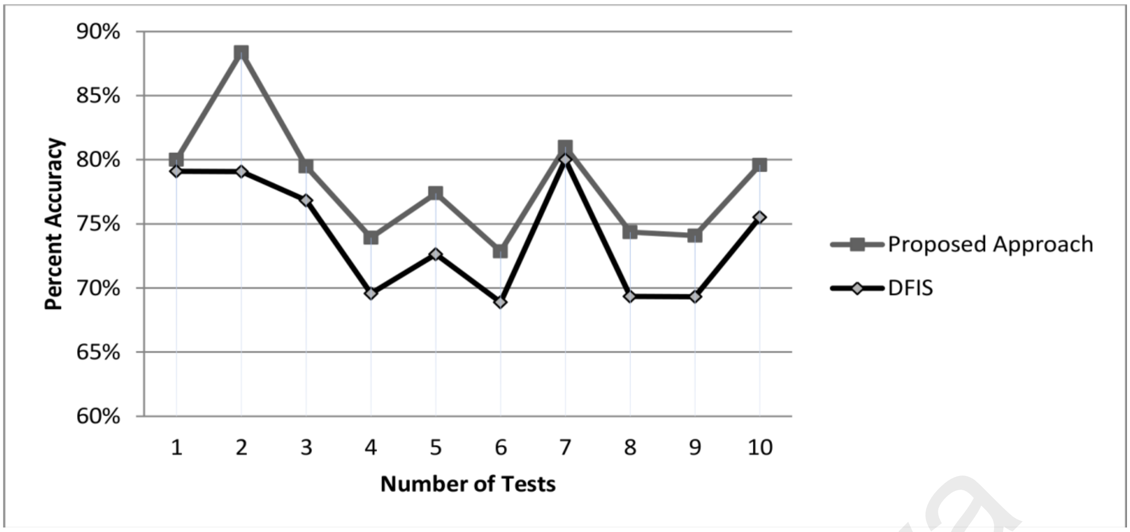


Figure 4.5: Prediction Accuracy Percentage of Flags Data Set

(c) SPECT Hearts Data Set

SPECT hearts is training data set containing images of SPECT abbreviated from Single Proton Emission Computed Tomography. The data base consists of 80 patients with 22 Boolean valued attributes. Numbers of values randomly deleted are 32, 98, 450, 182, 230, 62, 161, 47, 290 and 102. Percent performance graph is shown in Figure 4.6.

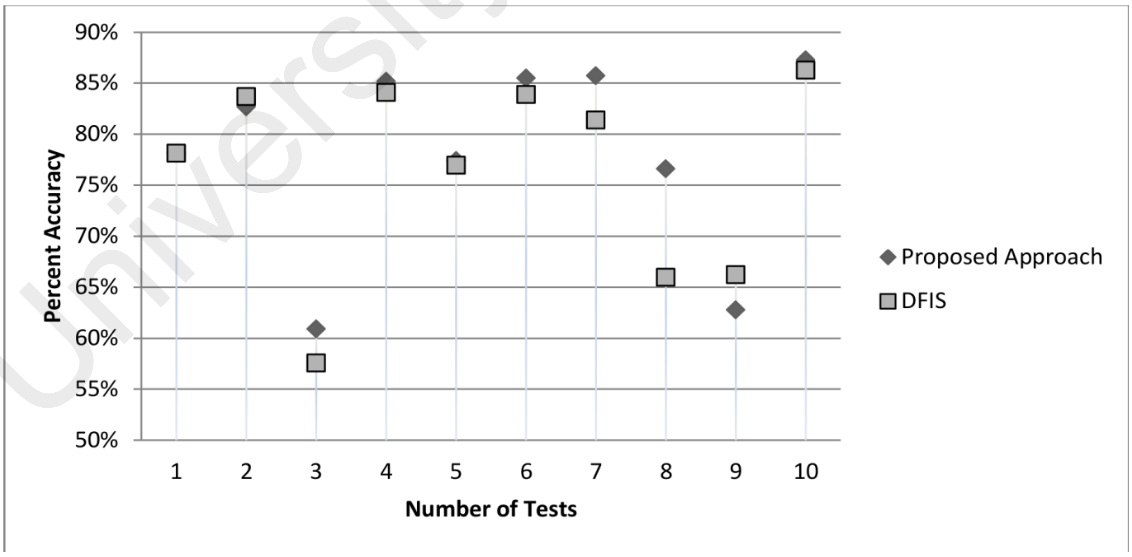


Figure 4.6: Percentage of accuracy graph of SPECT Hearts Dataset

Average accuracy of DFIS is 76.41% while that of proposed method is 78.20%. Hence, proposed method performs 1.80% better than DFIS for SPECT hearts data set.

(d) ***Congressional Votes Data Set***

This data set contains voting record of US congress members of 1984. 435 members had contested their votes in yes or no regarding 16 issues out of which only 230 member's votes are completed. This study selected these completed votes only for testing purpose and deleted randomly 161, 435, 122, 98, 263, 239, 205, 291, 424 and 136 values from this data set. After recalculating it though both approaches it was found that DFIS average accuracy is 65.50% while proposed approach has 72.98% accuracy.

Average performance of proposed approach is 7.84% better than DFIS for this data set. Performance graph of proposed approach vs. DFIS is plotted in Figure 4.7.

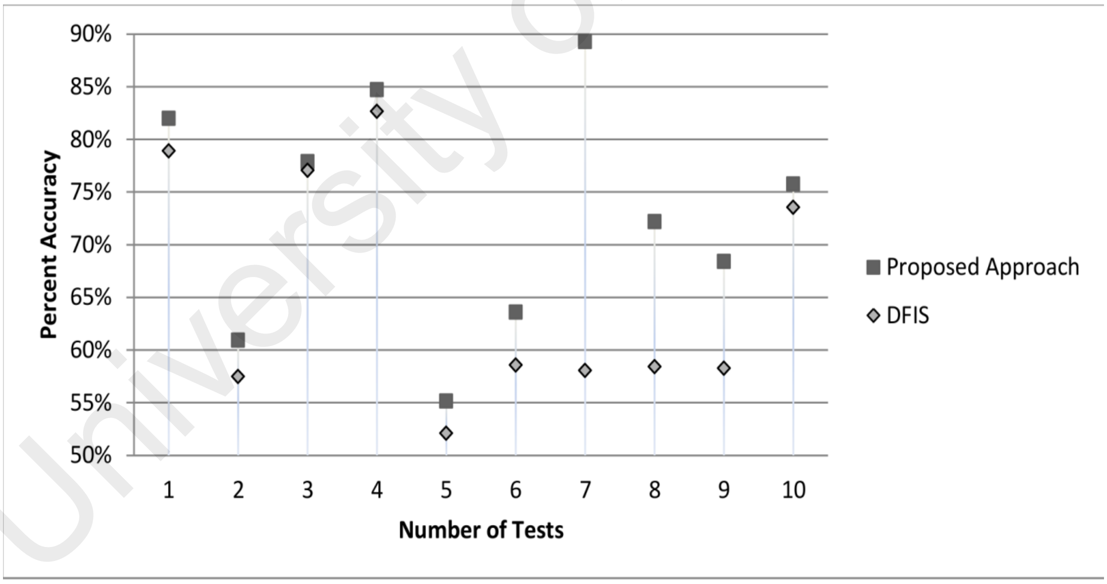


Figure 4.7: Percent accuracy graph of Congressional Votes data set

4.3.2.3 Causality Workbench LUCAP2 data set

Lung Cancer set with Probes (LUCAP) is an online data set containing Boolean valued artificially generated data by causal Bayesian networks. There are ten thousand imaginary objects (patients) with 143 features (symptoms) like Coughing, Fatigue, Yellow Fingers, Anxiety, Allergy, Attention Disorder and Smoking. This study selected first 1000 with all 143 parameters for its testing purpose. 322, 2354, 1190, 2083, 1432, 1158, 5413, 2457, 899 and 760 number of values are randomly deleted and recalculated it through DFIS and proposed method. It was found that for 1807 average unknowns, DFIS calculated 1294, while proposed method calculated 1328 accurate values. Hence, the average performance of proposed method is 1.89% better than DFIS for this data set. Percent accuracy graph of DFIS vs. proposed approach for LUCP2 data set is given in Figure 4.8.

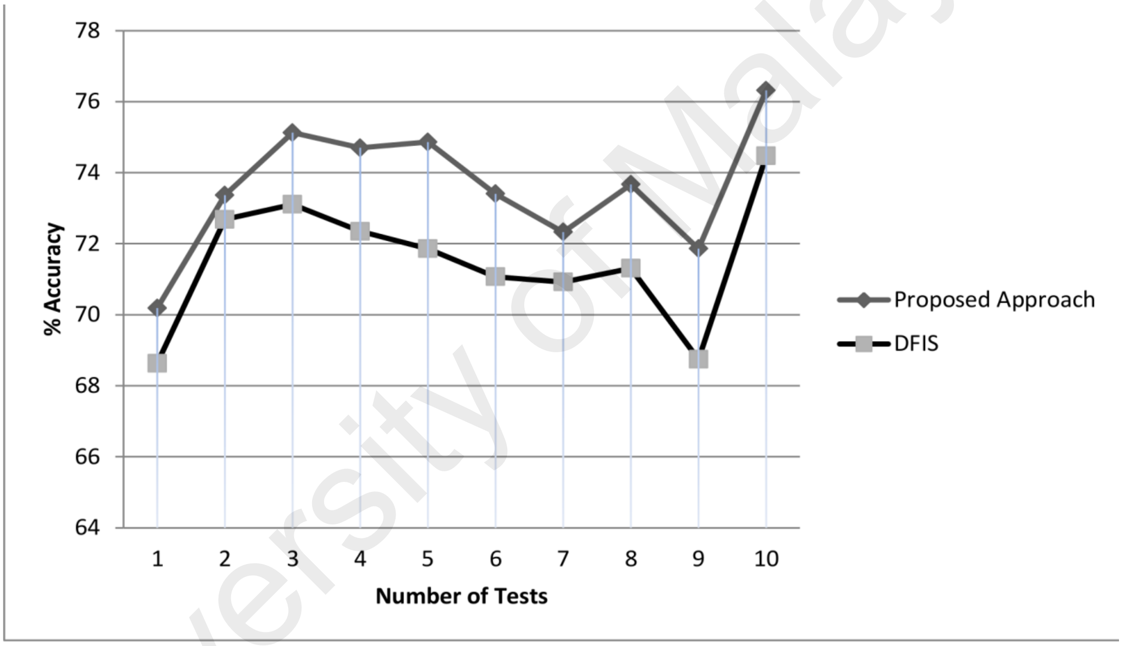


Figure 4.8: percent accuracy graph of LUCAP2 Dataset

4.3.2.4 Conclusion of overall results

In summary, the overall comparison results are given in the following Table 4.16.

Table 4.16: Comparison summary of all results

Data Sets	DFIS	Proposed approach	Improvement
Example 2	75.00%	83.00%	8.00%
Zoo Data Set	81.26%	84.67%	3.41%
Flags Data Set	74.02%	78.10%	4.08%
SPECT Hearts Data Set	76.41%	78.20%	1.79%
Congressional Votes Data Set	65.50%	72.98%	7.48%
LUCAP2 Data Set	71.61%	73.49%	1.89%

From Table 4.16, it is concluded that the proposed method performs better as compared to DFIS for each data set.

4.3.3 Discussions

In this section, some important queries that are raised regarding the threshold lambda (λ), its function, range and suitable values are discussed. The precise theoretical difference between DFIS and proposed, validation of proposed method and performance evaluation are also discussed. The question that "why UCI benchmark and LUCAP data sets are used? is answered, and significance of improvement in accuracy results is discussed.

The threshold lambda (λ) is a filter that can be set according to the requirements of individuals in getting weak or strong associations. Closer the value of λ to 1 result in more reliable association and closer the value to zero might result in selecting weaker associations. To select more than 50% associational results, the lambda must be fixed to 0.5 or above. In the incomplete case of Example 2.7 the threshold is kept as λ=0.85 to select only the parameters associations having minimum 85% similarity between them. The unknowns of parameters having less than 85% similarity are calculated through probability in DFIS while one of them ($*_1$) enters to the threshold range in the case of proposed approach. This reveals the core difference between DFIS and proposed approach. DFIS calculates all associations once for whole data set and assigns missing values according to it. It can be noticed that those parameters satisfying the threshold can be further categorized in less and more stronger association in the range between threshold and 1. Two parameters might have marginal similarity of 85% while another set of two may have stronger similarity as 90% or even 100%. DFIS treat them all as same for finding missing values, while proposed approach calculates the unknown first through the strongest among them and utilizes it for its role on upcoming calculations. This way, some of the unknowns that are calculated through probability enters association range and get more probable accurate results, as calculating unknowns through association is more reliable than probability (H. Qin, X. Ma, et al., 2012b). The results of DFIS are validated by calculating its decision values and comparing its MAPE with that of Zou et al. approach. As Zou et al. approach does not calculate missing values; therefore DFIS used indirect method of validation. But in this case, both DFIS and proposed approach calculate actual missing values and it is not needed to validate it through indirect decision values. So, this study uses direct method of comparing both techniques' actual results with original and the more accuracy of proposed approach validates its better performance.

This research takes DFIS as a benchmark study and DFIS has tested their algorithm on UCI benchmark data sets, therefore this research also use same data sets for it validation purpose. In addition to UCI benchmark data sets; this study also checks its accuracy on LUCAP data set which has artificially generated features.

From the results (Table 4.16), it can be observed that average accuracy for each data set varies from 1.8 to 4.1 percent better than DFIS. As missing values are predicted which are very important for its use in any application like decision making, therefore due to its importance its least accuracy is also considered as significant. For example the least accuracy is 1.79 and 1.89 percent better than DFIS for SPECT heart and LUCAP data sets. Both of these data sets are used for medical purposes and decision makings, therefore, even 1.79 and 1.89 percent improvement is an obvious significant values compare to DFIS.

4.3.4 Weaknesses of proposed work

Apart from improved accuracy, there are two main limitations of proposed approach as mentioned below.

4.3.4.1 Incorrect results rare cases

Sometimes the strongest association becomes false because of too much missing values or no real association existence. In this case, if missing values calculated in first step of proposed approach are incorrect then it affects the result of calculated values in next steps as well. This case can be viewed in the 2nd and 9th test result of SPECT Hearts data set graph where DFIS has high accuracy than proposed approach.

4.3.4.2 High computational complexity

High computational complexity of proposed approach compare to DFIS is obvious. DFIS accesses a data set of $m\times n$ size once for finding association while proposed approach accesses it $(m \times n)^2$ times during its execution. Complexity of proposed approach is DFIS times more than that of DFIS.

4.4 Conclusion

In this chapter, three previous approaches for prediction of incomplete soft set are discussed and DFIS is pointed out as the most suitable among them. An alternative approach of data filling for incomplete soft set is presented for the purpose of accuracy improvement. The process of DFIS is re arranged; therefore the maximum possible number of unknowns in incomplete soft set can be predicted through association between parameters. A modified algorithm is presented and proposed technique is explained with the help of an example as a proof of concept. The results of proposed method are compared with the existing approach (DFIS) after implementing both in MATLAB for four UCI benchmark data sets and Causality workbench lung cancer data set (LUCAP2) and shared the average results of both approaches in the form of graphs. Proposed approach has improved the accuracy of predicted unknowns significantly as compared to DFIS for all 5 data sets. Two main snags of proposed work are mentioned i.e. rare cases wrong values prediction and high computational complexity which can be resolved in its future work[6].

[6] These shortcomings of proposed work are avoided through clustering in the application of this method in next chapter of this thesis.

CHAPTER 5: APPLICATION OF DATA PREDICTION THROUGH STRONGEST ASSOCIATION IN ONLINE SOCIAL NETWORKS[7]

5.1 Introduction

Online social networks (OSNs) comprise three main elements: content, Web 2.0 technologies, and user communities (Ahlqvist, Bäck, Halonen, & Heinonen, 2008). Millions of people use OSNs to interact with one another, create content, share information, and exchange ideas in the virtual world. The data available in OSNs can provide researchers with insights into social networks and societies; these insights have been previously unattainable in both scale and extent (Lauw, Shafer, Agrawal, & Ntoulas, 2010). The interactions among users channeled through these OSNs create a huge amount of data, which are called user-generated data or social data. Social data constitute an immense source of information that spreads within each community on a global scale and reaches users, regardless of their status or location. The spread of information plays an important role in introducing new brands, promoting certain products, and achieving political goals by endorsing desired news and views (B. Min, Liljeros, & Makse, 2015). The information generated by every user is not necessarily spread efficiently in OSNs; only the information generated or promoted by specific eminent users, whose followers spread it on a large scale, is spread efficiently. Such users have either already gained celebrity status before connecting to social media or achieved that status on social media because of their fascinating social activities and involvement with other members. PageRank, *k*-core, and centrality algorithms are used to identify these top spreaders. After being identified, the top spreaders can be handled optimistically, blocked from spreading unwanted content, or leveraged to accelerate the

[7] The main idea of this chapter has already been published in ISI indexed journal “IEEE Access” with the title “Virtual Community Detection through the Association between Prime Nodes in Online Social Networks and Its Application to Ranking Algorithms”.

spread of positive or desired information. Numerous efforts have been exerted to identify top users (Brin & Page, 2012; Duanbing Chen, Lü, Shang, Zhang, & Zhou, 2012; De Domenico et al., 2013; Liu, Tang, Zhou, & Do, 2015; B. Min et al., 2015; Morone & Makse, 2015; S. Pei et al., 2014; S. Pei, Muchnik, Tang, Zheng, & Makse, 2015); however, this problem has remained unsolved, mainly because not all the connections in an OSN can be completely collected given that most OSNs impose certain privacy and technical restrictions. Consequently, incomplete network data may reduce the accuracy of ranking algorithms (B. Min et al., 2015; S. Pei et al., 2014). Therefore, this study proposes a method for completing incomplete OSNs to a reliable degree before applying ranking algorithms. OSN completion using any suitable link prediction technique can help improve the accuracy of ranking algorithms. This study contributes to the existing literature by introducing a novel method for OSN.

Researchers have attempted to detect network communities (Bedi & Sharma, 2016; Fortunato, 2010; Palla et al., 2007; Peng et al., 2014; Radicchi et al., 2004; Sun, 2016; Zhan et al., 2016), proposed various definitions, and concluded that "its elements are highly interconnected" (Güneş et al., 2016). Progress has been achieved in terms of completing an incomplete network (i.e., an OSN) by predicting new links (Adamic & Adar, 2003; Duan et al., 2016; Güneş et al., 2016; Kossinets, 2006; D. Li et al., 2016; Liben-Nowell & Kleinberg, 2007; Lü & Zhou, 2011; Newman, 2001). Link prediction is divided into two categories: network topology based and node based (Güneş et al., 2016). Link prediction approaches that use network topology are based on the fact that communities utilize different aspects of common neighbors but their main focus is on "interconnection among nodes" with its own significance (Güneş et al., 2016; Zhan et al., 2016). The current study proposes the hypothesis that maximum nodes inside an OSN belong to different virtual communities, and a community member exhibits a behavior similar to that of other community members, particularly in terms of linking to prime nodes. Community formation may be induced by the direct physical and real similarities among members based on their geography, locality, and occupation as well as real-world themes in society. It may also be induced by indirect odd and virtual connections based on similar personal choices, cognitive levels, acceptance and rejection behavior, and ideology, regardless of physical, real, and geographical interactions. The identification of virtual communities in OSNs can provide researchers more insights to stimulate further discussion, inspire new ideas, and lead to alternative conclusions. In this study, virtual communities are identified and used in missing link prediction. The identified communities and predicted links are applied to improve the accuracy of existing ranking algorithms and proposed for the future growth of OSNs. A virtual community with four nodes, which represent the community members, and two prime nodes, which represent the common interests of the community members, is shown in Fig. 1. Nodes *b*, *c*, *d*, and *e* are similar and form one virtual community by connecting to their prime node *f*. The same community is also connected to their second prime node *a* except *c*, which according to proposed approach, should also connect to *a* while a connection between *a* and *f* exists based on the consensual definitions of the common neighbor approaches. The main differences between the definitions of previous approaches and the proposed approach are presented in Table 5.1 by using the example in Figure. 5.1.

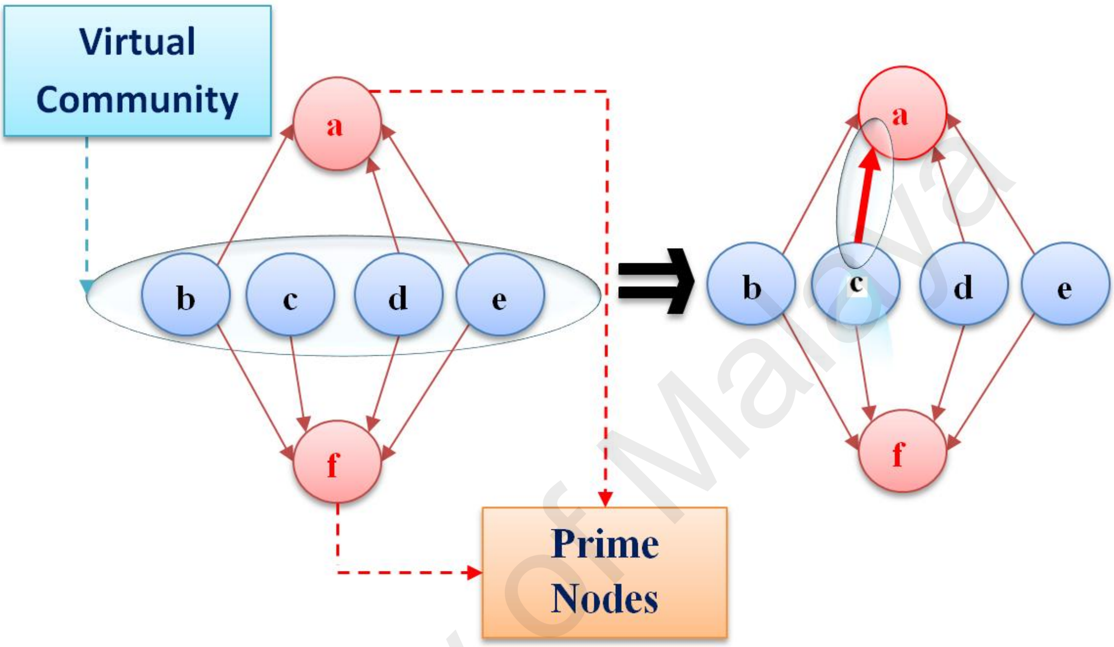


Figure 5.1: Graphical description of a virtual community with members *b*, *c*, *d*, and *e* and its nodes of interest (prime nodes) *f* and *a*. The highlighted link from *c* to *a* indicates that *c* should be connected to *a* to behave like other community members.

Table 5.1: Differences between the proposed approach and existing approaches to community detection and link prediction

Difference	Consensual definitions of existing community and common neighbor approaches	Proposed approach to detect a virtual community and predict the links in the community
1	Nodes *a* and *f* have common neighbors; therefore, *a* and *f* belong to the same community.	Nodes *b*, *c*, *d*, and *e* are similarly linked to prime node *f*; therefore, *b*, *c*, *d*, and *e* form a virtual community.
2	Common neighbors (*b*, *d*, and *e*) are mandatory for community (*f* and *a*) formation.	No common neighbor is considered in virtual community (*b*, *c*, *d*, and *e*) formation, although community members may be interconnected.
3	A new direct link between nodes *f* and *a* should exist because both belong to the same community and have common neighbors.	Nodes *f* and *a* are prime nodes of a virtual community. No direct link is established between them until either of these nodes becomes a prime node and the other enters the virtual community of the prime node in any of the succeeding iterations.
4	Node *c* has no common neighbor with *a*, therefore a new link from *c* to *a* is not established.	Node *c* is part of a virtual community (*b*, *c*, *d*, and *e*) and completely connected to prime node *f*. All community members except *c* (*b*, *d*, and *e*) are connected to another prime node *a*; therefore, node *c* should also connect to *a*.

The data-filling approach for an incomplete soft set (DFIS) (H. Qin et al., 2012a) and PSA (prediction through strongest association, proposed in the previous chapter) use the association between parameters to predict missing data in an incomplete soft set. Inspired by DFIS and PSA, the proposed OSN completion method uses the association between nodes in OSN community detection and link prediction. Aside from community detection, the current work includes the link prediction technique developed by Li et al. (D. Li et al., 2016). Li et al. approach relies on link prediction through information diffusion but disregard the community association factor. By contrast, the proposed method identifies the main probable reason for information diffusion and applies these methods without going into the details of diffusion. The differences and similarities of Li et al. method and the proposed method are further discussed in the related sections of discussion (section 5.4). The results of the proposed approach are validated using the ranking algorithm list obtained by tracking diffusion links under the real spreading dynamics of information (S. Pei et al., 2014).

The main contributions of this chapter are as follows:

a. The virtual communities in OSNs are indentified whose elements exhibit similar behavior in linking to their nodes of interest (prime nodes)
b. New links in incomplete OSNs are predicted up to the degree of strong association between its prime nodes through virtual communities.
c. The results of the proposed method are validated by applying two well-known ranking algorithms, namely, PageRank and *k*-Core, to real and large data sets, which are extracted from Facebook and Twitter, and subsequently compare their ranking accuracy rates before and after OSN completion.

In addition to above contributions, the validation part expands the practical application of proposed OSN completion to the improvement in the accuracy of ranking algorithms.

5.2 Rudimentary Concepts

This section discusses the background of incomplete data completion by prediction through the association between parameters through PSA and DFIS (H. Qin, X. Ma, et al., 2012a) and the improvement of existing ranking algorithms (Bakshy, Hofman, Mason, & Watts, 2011; Cha, Haddadi, Benevenuto, & Gummadi, 2010; W. Chen, Cheng, He, & Jiang, 2012; Jabeur, Tamine, & Boughanem, 2012; Kim & Han, 2009; Lü, Zhang, Yeung, & Zhou, 2011; Romero, Galuba, Asur, & Huberman, 2011; Silva, Guimarães, Meira Jr, & Zaki, 2013; Weng, Lim, Jiang, & He, 2010).

5.2.1 Incomplete Data Completion by Prediction through the Association between Parameters

An initial attempt to calculate the decision values in an incomplete soft set was made by applying the weighted average method (Zou & Xiao, 2008). Recently, however, the same decision values were obtained, and certain rational values were simultaneously assigned to missed values by applying a less complex method of using probabilities for 1s and 0s (Kong et al., 2014). In both weighted average (Zou & Xiao, 2008) and probability (Kong et al., 2014) methods, the integrity of the standard soft set is damaged, and the set is converted into a fuzzy soft set. DFIS (H. Qin, et al., 2012a) and PSA prioritizes the prediction of missing data in a soft set through the association between parameters and assigns second priority to probabilities. A soft set is a mathematical tool for efficiently handling uncertain or vague data (Molodtsov, 1999); however, the association between parameters is not limited to uncertain data given that it also applies to actual daily life data. The association between parameters can be illustrated in the following example.

Suppose four candidates are under consideration based on four parameters, as presented in the BIS provided in Table 5.2. A parameter that belongs to a candidate is represented by 1; otherwise, 0. The parameters “young” and “having children” have an inconsistent association with each other, i.e., a young candidate is more probable to be unmarried and have no children, and vice versa. By contrast, a consistent association exists between the parameters “young” and “inexperienced”, i.e., a young candidate is more likely to be inexperienced, and vice versa.

Table 5.2: Representation of candidate's file (BIS)

Candidate/Parameter	Young	With children	Highest degree is PhD	Inexperienced
C_1	1	0	1	1
C_2	0	1	0	0
C_3	0	0	0	0
C_4	1	0	0	0

In daily life, association may be a logical probability for general cases but may not hold true for every case. As shown in Table 5.2, C_1 and C_4 are young and have no children, C_2 is old and has children, and C_3 is also old but has no children. In general, the association (consistent or inconsistent) results in over 50% accuracy because general cases are more than specific cases in all instances. This study extends this association to the prediction of missing OSN nodes.

5.2.2 Improvement of Ranking Algorithms for OSNs

Numerous studies (Bakshy et al., 2011; Cha et al., 2010; W. Chen et al., 2012; Jabeur et al., 2012; Kim & Han, 2009; Lü et al., 2011; Romero et al., 2011; Silva et al., 2013; Weng et al., 2010) have contributed to the improvement of existing ranking algorithms by introducing new factors into these algorithms. However, most current studies have used only partial data from OSNs to test their proposed ranking algorithms. Presumably, this study is the first to propose the prediction of missing nodes to improve ranking algorithms. The prediction of some missing nodes that cannot be extracted from OSNs because of the restrictions imposed by these OSNs is assumed capable of improving the accuracy of ranking algorithms using partial network data.

5.3 Materials and Methods

In this section, the prediction algorithm, ranking algorithms, and data sets used in the study are discussed. The validation of the proposed approach is also presented.

5.3.1 Prime Node Association in an OSN and Completion of an Incomplete OSN

If a small network includes users *a, b, c, d, e,* and *f,* with *a* following all the other users in this group and *b* following all the other users except *e*, and the group is known to contain missing values, then *b* most probably also follows *e* because *b* is similar to *a* in certain characteristics (i.e., following users in the group). Thus, a consistent association exists between *a* and *b*. The steps of the method for predicting missing nodes in an OSN through association are explained in the following subsections, with each step being illustrated by an example as a proof of concept.

5.3.2 Representation of an OSN as a BIS

For association determination, an OSN should be converted into a BIS. An OSN link consists of two types of nodes: "followee" and "follower" or "linked." Unique sets of nodes are selected from both the followees and followers in the group being considered and are then represented in rows and columns. A cell *x* with index *ij* is assigned a value equal to 1 if node *i* is connected to node *j*, i.e., $x_{ij} = 1$; otherwise, $x_{ij} = 0$. Definitions 5.1 and 5.2 explain the connection between two nodes.

Definition 5.1: For two nodes *x* and *y*, if *x* is following *y*, then they are represented by *xy* wherein x is connected to y.

Definition 5.2: For two nodes *x* and *y*, if *x* is not following y, then they are represented by *x**y* wherein *x* is not connected to *y*.

Definitions 5.3 and 5.4 determine the linked nodes.

Definition 5.3: Nodes in the combination $x_i y_i$, which is represented by 1 in BIS, are called linked nodes.

Definition 5.4: Nodes in the combination $x_i * y_i$, which is represented by 0 because $i \neq j$ in BIS, are called unlinked nodes.

The following example clarifies the concept of linked nodes based on Definitions 5.1–5.4.

Example 5.1: A group has six nodes, i.e., *a, b, c, d, e,* and *f*. The links of the nodes are *ab, ac, ad, ae, ba, bc, bd, bf, be, cd, cf, da, dc, df, ea, ed, ef,* and *fe*. The unique nodes of both followees and followers are *a, b, c, d, e,* and *f*. In Table 5.3, both columns and rows represent all the nodes in the group (i.e. small OSN).

Table 5.3: Representation of the OSN as a BIS

Followee/Follower	*a*	*b*	*c*	*d*	*e*	*f*
a	0	1	1	1	1	0
b	1	0	1	1	0	1
c	0	0	0	1	0	1
d	1	0	1	0	0	1
e	1	0	0	1	0	1
f	0	0	0	0	1	0

As shown in Table 5.3, *a* is connected to *b, c, d,* and *e*; therefore, $x_{12} = x_{13} = x_{14} = x_{15} = 1$. By contrast, *a* is not connected to *a* and *f* (*a*a* and *a*f*), therefore $x_{11} = x_{16} = 0$. All the other cells are assigned values using the same method.

5.3.3 Incomplete OSN

Some links in the BIS-represented OSN under consideration are supposed to be missing because of user privacy or OSN-imposed restrictions. If a node *x* is linked to *n* number of nodes, then its values are represented by 1 in the corresponding *n* cells of the BIS. Evidently, these *n* links are definitely the mandatory part of the OSN, and these values cannot be changed during network completion through links prediction. However, other nodes of number *m* may not follow the same node *x*; therefore, each of these nodes and node *x* are unlinked nodes and the values of node *x* are represented by 0 in the corresponding cells of the BIS. Some or all of these *m* number of unlinked nodes can be added only to the followers of *x* to complete the network. That is, only the cells of the BIS with a value equal to 0 can be considered for connecting to node *x*. If the link prediction technique identifies such nodes to be the followers of node *x*, then the corresponding value of *x* in the BIS will be changed from 0 to 1. This process is further explained in the following example.

Example 5.2: The same group in Example 5.1 is used in this example. Suppose this small network is incomplete. In this case, *a* is followed by all the other nodes in the group except *c*, *f*, and *a* itself. Node *a* cannot follow itself (condition $i = j$ of Definition 5.4); therefore, *a* may be followed by *c* and *f* to create a complete network. These probable missing links $c * a$ and $f * a$ are represented by * in Table 5.4 and targeted for prediction through association with other followee nodes.

Table 5.4: Representation of an incomplete partial OSN as a BIS

Followee/Follower	*a*	*b*	*c*	*d*	*e*	*f*
a	0	1	1	1	1	0
b	1	0	1	1	0	1
c	*	0	0	1	0	1
d	1	0	1	0	0	1
e	1	0	0	1	0	1
f	*	0	0	0	1	0

Similarly, other cells with $x_{ij} = 0$, where $i \neq j$, can be considered unknown and targeted for prediction through association column wise.

5.3.4 Prediction of unknown links through association

Unknown links can be predicted through association by regarding the unlinked nodes of the first column of the BIS as unknown, i.e.,

$$\text{If } x_{i1} = 0, \text{ then } x_{i1} = * \text{ for } i \neq j. \qquad (5.1)$$

In the next step, the consistency of the first column with all the other columns is calculated using

$$CN_{1k} = \{x \mid F(x_{i1}) = F(x_{ik})\}, \tag{5.2}$$

where $F(x_{i1})$ denotes all the cell values in the first column of the BIS, $F(x_{ik})$ denotes all the cell values in all the other BIS columns, and CN_{1k} is the set of cells in the first column that are consistent with the correspondent cells in the other *k* columns.

The consistency degree is calculated using

$$CD_{1k} = \max\left(\frac{CN_{1k}}{U_1}\right), \tag{5.3}$$

where CD_{1k} is the highest consistency ratio of the first column with *k*-th column and U_1 is the number of known values in the first column. The latter is calculated using $U_1 = |\{x \mid F(x) = 1\}|$, which indicates that U_1 is the number of 1s or known values in the first column. The threshold range $1 \geq \lambda > 0$ is a predefined filter to select strong associations.

Definition 5.5: Column a is consistent with column b, that is, $a \Leftrightarrow b$, if $CD_{ab} \geq \lambda$.

If $CD_{1k} \geq \lambda$, then the unknown values in the first column are calculated as the corresponding values of the k-th column. Thus, if $1 \Leftrightarrow k$, then the unknown values $x_{i1} = *$ are

$$x_{i1} = x_{ik}, \tag{5.4}$$

where x_{i1} denotes the unknown values in column 1, and x_{ik} denotes their corresponding values in the k-th column.

The following definitions of “prime nodes” and “virtual community” are derived from the illustration of Definition 5.5.

5.3.4.1 Prime nodes

Prime nodes are the nodes that represent two or more consistent columns. For example, columns 1 and 6 in Table 5.4 are consistent with each other, and they are represented by nodes a and f, respectively; therefore, nodes a and f are called prime nodes.

5.3.4.2 Virtual community

Virtual community is the union set of the followers of the prime nodes. For example, nodes *a* and *f* form a set of prime nodes. In this set, the followers of *a* are *b*, *d*, and *e* (Table 5.3), and the followers of *f* are *b*, *c*, *d*, and *e*; thus, the union set of the followers of prime nodes *a* and *f* comprises *b*, *c*, *d*, and *e*, which form the virtual community with respect to prime nodes *a* and *f*

Definition 5.6: Column a is non-consistent with column b, that is, $a \nLeftrightarrow b$, if $CD_{ab} < \lambda$

If $CD_{1k} < \lambda$, then the unknown values are reverted to their original value of 0. Thus, if $1 \nLeftrightarrow k$, then the unknown values $x_{i1} = *$ are

$$x_{i1} = 0 \tag{5.5}$$

After values have been assigned to the unknown values in the first column, the processes of assigning unknown values, calculating consistency and its degree, and predicting unknown values using Equations (5.1) - (5.5) are repeated individually for all the other columns. The updated links are obtained from the updated BIS, which may consist of more links than the original links if the new links are predicted through association. The following example illustrates the process of calculating column consistency.

Example 5.3: In this example, the incomplete sample OSN in Example 5.2 has been completed through association between nodes. Table 5.4 presents the incomplete OSN group with the unlinked nodes of the first columns assigned to be unknown values based on Equation (5.1).

The consistency of the first column with the second column can be determined using Equation (5.2). No corresponding elements in columns 1 and 2 are the same; therefore, $CN_{12} = 0$. For column 3, only its second and fourth corresponding elements are the same as those in column 1; therefore, $CN_{13} = 2$. Similarly, $CN_{14} = 2$, $CN_{14} = 0$, $CN_{15} = 0$, and $CN_{15} = 3$. Three 1s are present in the first column; therefore, $U_1 = 3$. From Equation (5.3),

$$CD_{1k} = \max\left\langle \frac{CN_{12}}{U_1}, \frac{CN_{13}}{U_1} \frac{CN_{14}}{U_1} \frac{CN_{15}}{U_1} \frac{CN_{16}}{U_1} \right\rangle = \max\left\langle \frac{0}{3}, \frac{2}{3}, \frac{2}{3}, \frac{0}{3}, \frac{0}{3}, \frac{3}{3} \right\rangle = 1 \Rightarrow CD_{16} = 1$$

If the threshold is $\lambda = 0.8$, that is, $CD_{16} = 1 > \lambda$; then column 1 is consistent with column 6.

From Definition 5.5 as $1 \Leftrightarrow 6$ for $x_{31} = x_{61} = *$ therefore, from Equation (5.4), $x_{31} = x_{36} = 1$ and $x_{61} = x_{66} = 0$, as highlighted in Table 5.5.

Table 5.5: Representation of an incomplete OSN after partial completion using association between nodes

Followee/ Follower	*a*	*b*	*c*	*d*	*e*	*f*
a	0	1	1	1	1	0
b	1	0	1	1	0	1
c	**1**	0	0	1	0	1
d	1	0	1	0	0	1
e	1	0	0	1	0	1
f	**0**	0	0	0	1	0

A comparison of the BIS in Table 5.5 with the initial BIS in Table 5.5 shows that the new link *ca* has been created during the first phase of prediction through the association between the nodes. The unknown values can be predicted using the same method for the second, third, fourth, fifth, and sixth columns.

The algorithm for the prediction of missing nodes through association is presented as given in Figure 5.2.

```
Prediction of missing nodes

Input: OSN clusters with missing nodes

Output: Complete OSN

1. Convert OSN clusters into a BIS.
2. Assign j pointer to the first column.
3. Render the unlinked nodes in the j-th column unknown
4. Calculate the consistency of the j-th column with all the other columns k ($CN_{jk}$).
5. Divide all the consistency values by the number of linked nodes ($U_j$) of the j-th column and find its maximum $CD_{jk}$.
6. If $CD_{jk} \geq \lambda$, then the unknown values in the j-th column are the same as the corresponding values in the k-th column; otherwise, the unknowns are 0.
7. Increase the j-th counter by 1 until the last column, and return to step 3.
8. Convert the BIS into OSN clusters and combine the OSN clusters.
9. End
```

Figure 5.2: Algorithm for the prediction of missing nodes

5.3.5 Ranking Algorithm

Researchers have proposed various algorithms to detect and rank top spreaders in OSNs. Among these, PageRank and *k*-core are considered the most outstanding and widely used algorithms.

5.3.5.1 PageRank

PageRank is a network-based diffusion algorithm originally proposed by Brin et al. (Brin & Page, 2012). This well-known algorithm is used by the Google search engine for ranking web pages. It allows for the global ranking of all web pages based only on their connected links and locations in the web graph, regardless of their content. PageRank calculates recursively and considers two main parameters, namely, the number of inbound links and their corresponding PageRank values.

5.3.5.2 k-Core ranking

In *k*-core-based ranking, each node is assigned a *k*-shell number k_s, which is the order of the shell to which it belongs. Initially, the *k*-shell eliminates all the nodes with a degree (*k*) of 1. The elimination process continues until all the nodes with a degree of 1 are eliminated. Similarly, this elimination procedure is applied to the next *k*-shells. This decomposition process is repeated until the *k*-core of the network is detected (Batagelj & Zaversnik, 2003).

5.3.6 Data sets

The following real and large OSN data sets are used in this study.

5.3.6.1 Facebook data set

This social network data set contains 63,520 nodes and 1,545,686 edges. Its wall post data set consists of 876,993 wall posts from 46,952 users. This data set was used in a recent study of Pei et al.(S. Pei et al., 2014).

5.3.6.2 Twitter data set

The Twitter data set (De Domenico et al., 2013) is the Higgs data set constructed before, during, and after the announcement of the discovery of a new particle with the features of the elusive Higgs boson on July 4, 2012. The data used constitute the data extracted from Twitter between July 1, 2012 and July 7, 2012. This social network data set contains 456,626 nodes and 14,855,842 edges. The retweet network contains 256,491 nodes and 328,132 edges. On the basis of these data sets, the social network nodes for this study are used to construct the network and retweet data sets for the diffusion graph.

5.3.6.3 Important features of the data sets

The Facebook data set has two elements, i.e., social network and wall posts. During the data extraction process, some links might have been lost because of the parameters specified by the extracting body, the privacy constraints implemented by the operators, and user options. In the validation of social network completion, wall posts represent actual spreading by the user. Thus, the predicted data must contain the links between *i* and *j* users. These links are present in the wall posts diffused by *i* from *j* and missing in the extracted social network sub-data set. This scenario is reflected in the results, and the statistics are presented in Section 4. Similarly, the Twitter data set has two elements, namely, social network and retweets.

5.3.7 Performance Evaluation

To evaluate the validity of the proposed link prediction method for OSN completion, this study uses the ranking algorithms PageRank and *k*-core to identify the top spreaders before and after the completion of both networks and subsequently compares the results. The spreading efficiency or influence $inf(i)$ of each user *i* is calculated as the number of users influenced by user *i* based on the wall post data set of Facebook and the retweet data set of Twitter. These influenced users are those who propagate the information of user *i*, and $inf(i)$ is obtained using breadth-first search for user *i* (S. Pei et al., 2014). Information spreading is in the form of sharing the wall posts of user *i* in Facebook and retweeting his or her tweets in Twitter. The retweet network serves as an illustrative network that explains how content is propagated (De Domenico et al., 2013). The variable $inf(i)$ is used to calculate the average spreading efficiencies M_{st} of the set of top spreaders under consideration. Sets of top spreaders may represent the top 1%, 5%, 10%, 20%, 30%, and 50%, and their average influence levels in wall posts and retweets are considered the standard M_{st}. Similarly, the average influence levels of the same set of top spreaders are calculated using the ranking algorithms for the network before prediction (M_{bp}) and after link prediction (M_{ap}). For the comparison of the accuracy rates of the ranking algorithms before and after network completion by link prediction, the imprecision functions ε_{bp} before link prediction and ε_{ap} after link prediction are used as proposed in (Kitsak et al., 2010) and given as:

$$\varepsilon_{bp} = 1 - \frac{M_{bp}}{M_{st}}, \tag{5.6}$$

$$\varepsilon_{ap} = 1 - \frac{M_{ap}}{M_{st}} \tag{5.7}$$

The lower the value of the imprecision function (ε), the more accurate the prediction, and vice versa. An ε value that is close to 0 denotes high efficiency because the selected nodes are the same as those that contribute the most to information diffusion.

5.4 Results and discussions

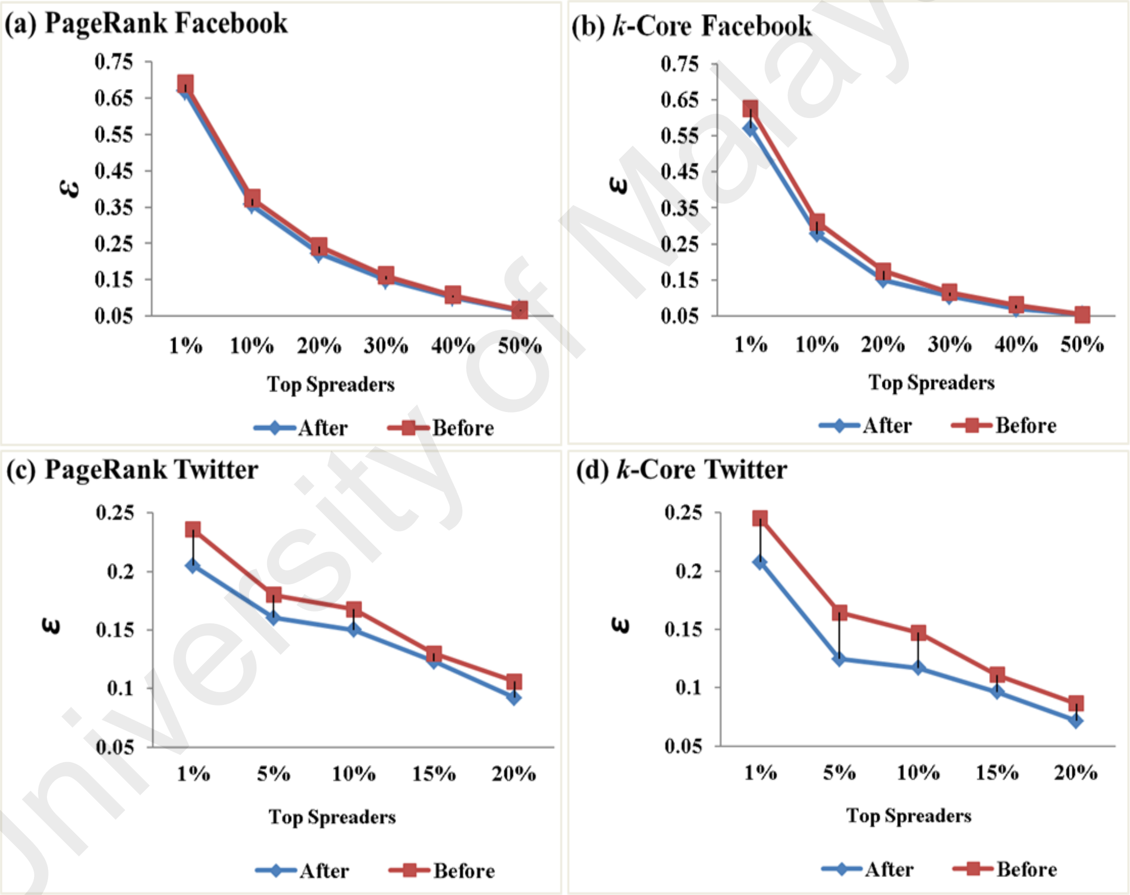
The obtained results are reported and discussed in this section. The statistics of both data sets after link prediction are given in Table 5.6.

Table 5.6: Statistics of the prediction results

Data Set	Number of Links Before Link Prediction	Number of Links After Link Prediction	Number of New Predicted Links	Percentage of New Predicted Links
Facebook	1,545,686	1,637,012	91,326	5.91%
Twitter	14,855,842	16,288,346	1,432,504	9.64%

The imprecision function values for the top 1%, 10%, 20%, 30%, and 50% top spreaders identified by PageRank and *k*-core for the Facebook data set are compared in

Figure 5.3(a) and 5.3(b) and their statistics are given in Table 5.6. For the Twitter data set, the imprecision function values for the top 1%, 5%, 10%, 15%, and 20% are compared in Figure 5.3(c) and 5.3(d) and their statistics are given in Table 5.8. The average imprecision function values before and after link prediction for both data sets and the two ranking algorithms are presented in Figure 5.3(e). Network samples created through Ghepi for 10 nodes before and after prediction for both data sets are given in Figure 5.4.



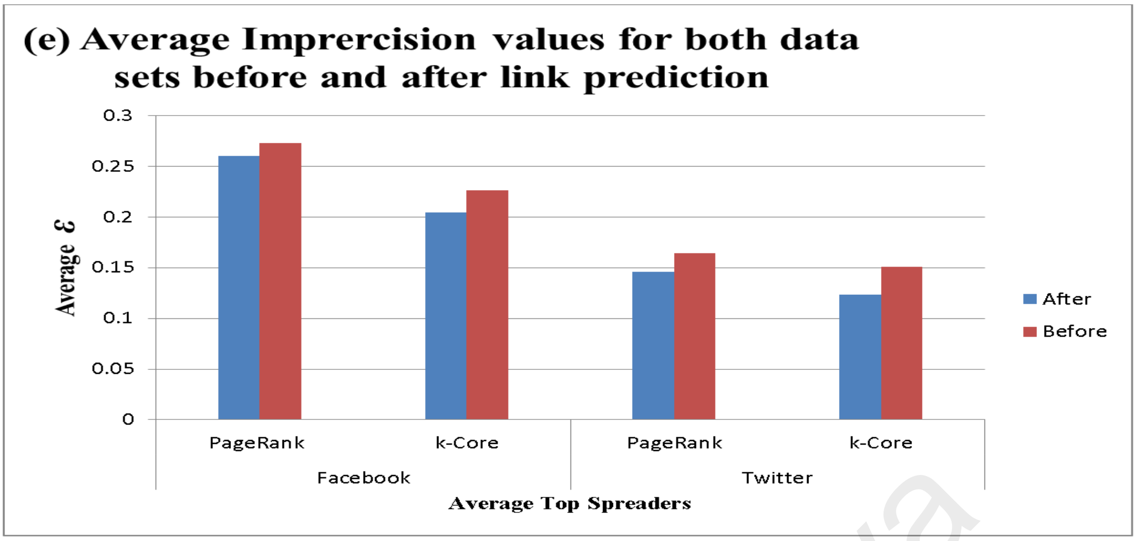


Figure 5.3: Accuracy improvement graphs using the imprecision function ε: (a) PageRank and (b) k-core for the Facebook data set; (c) PageRank and (d) k-core for the Twitter data set; (e) average of the results presented in (a), (b), (c), and (d).

Table 5.7: Statistics of imprecision for Facebook data set

	PageRank			k-Core		
Top Spreaders	After	Before	Improvement	After	Before	Improvement
1%	0.6652	0.6876	0.0224	0.571	0.6262	0.0552
10%	0.3526	0.3729	0.0203	0.2774	0.3109	0.0335
20%	0.2235	0.2427	0.0192	0.1502	0.1737	0.0235
30%	0.15	0.161	0.011	0.1043	0.1146	0.0103
40%	0.1016	0.1058	0.0042	0.07	0.0804	0.0104
50%	0.0658	0.0664	0.0006	0.053	0.0534	0.0004

Table 5.8: Statistics of imprecision for Twitter data set

	Imperceision for PageRank			Imprecision for k-Core		
Top Spreaders	After	Before	Improvement	After	Before	Improvement
1%	0.2049	0.2356	0.0307	0.2076	0.2451	0.0375
5%	0.1603	0.1801	0.0198	0.1247	0.1646	0.0399
10%	0.1503	0.1676	0.0173	0.1171	0.1471	0.03
15%	0.1234	0.1301	0.0067	0.09648	0.111	0.01452
20%	0.09195	0.1061	0.01415	0.0716	0.0864	0.0148

The improvement in the accuracy of the ranking algorithms after network completion can be explained logically. Prediction through the association between nodes includes only those nodes that demonstrate a behavior similar to those of other nodes in following the same node. Most probably, the predicted nodes are the actual followers. This premise is verified by the improvement in the accuracy of the ranking algorithms achieved during the validation phase. The question as to why nodes with similarities in following other nodes are more likely to have similar followees constitutes the core idea of proposed approach. In Example 5.3, nodes *b*, *c*, *d*, and *e* are following node *f*, and all the followers of user *f* except *c* are following *a*. The first probable answer to the core question may be obtained by determining the reason why user *c* follows *a* is that users *b*, *c*, *d*, and *e* appear to belong to the same community in real life. Nodes *a* and *f* have similar relationships with this community (*b*, *c*, *d*, and *e*); 100% of the community members is following *f*, and 75% of the community members is following *a*. User *c* may not be aware of the existence of *a* in social media but will follow *a* after coming to know him or her through any channel.

Physical community relationship can be established by either living in the same geographical area, sharing the same workplace, or being members of the same institution. The relationships to nodes *f* and *a* depend on the prime nodes. In this case, these prime nodes may represent prominent persons in their geographical locality, organizational team leaders, teachers, or any other possible relationship based on ground situation.

The second probable answer is that, even if these users (*b*, *c*, *d*, and *e*) share no such physical community relationship in real life, their preferences are correlated and important in an effective virtual relationship. This virtual or social association can be ascribed to similar choices or shared worldview, whereas their relationship with prime nodes *a* and *f* can be that of a certain product, intellectual, ideologist, or any other possible relation. As followers of node *f*, nodes *b*, *d*, and *e* may interact with the posts of the influencer by sharing, liking, or commenting. As a follower of user *d*, user *c* will find posts of *a* from *d* or other nodes and make him or her a probable followee. A set of related prime node links is not limited to one community. From existing common neighbor approaches to link prediction, a new link can be predicted between user *f* and *a* in this incomplete case. New link prediction between user *f* and *a* is sensible when the followers share the same physical community relationship with their followees. However, in the case of different brands products and choices, the existence of a link between *f* and *a* has low probability between two competitors. The method established by Li et al. (D. Li et al., 2016) differs from proposed approach because their approach assumes that *c* may follow *a* because of the posts of *a* being shared by *d*. However, they did not consider the probability of similarity between *c* and other community members; that is, if *c* receives no information diffused from *a* through his or her friend *d*, then *c* still tends to connect with *a* and can be suggested in the OSN recommender system. According to Li et al., *c* may also follow *e* because of the information reaching him or her via *d*. In proposed approach, the probability of a new link formation between *c* and *e* is less compared with that of similarity new link formation between *c* and *a*.

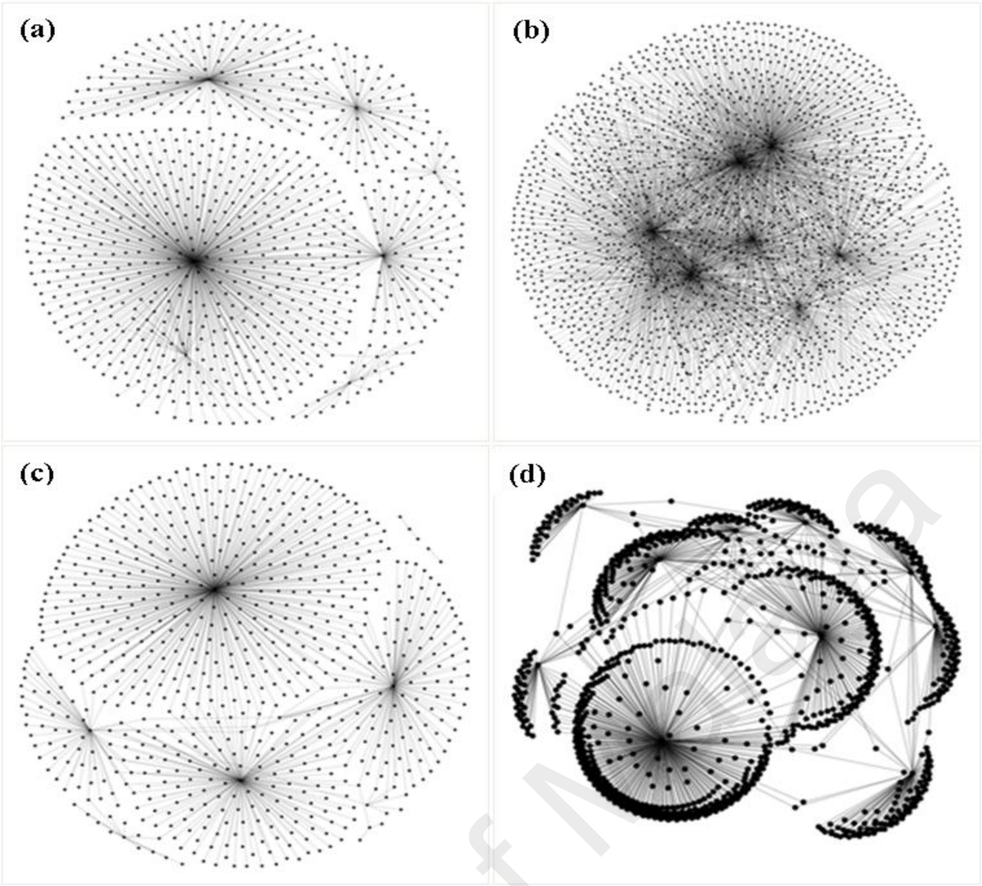


Figure 5.4: Samples created using Gephi for both data sets before and after link prediction: (a) 10 nodes of the Facebook data set before prediction and (b) the same 10 nodes after link prediction; (c) 10 nodes of the Twitter data set before prediction and (d) the same 10 nodes after link prediction.

The experimental results show that the prediction of the missing links between users has improved the ranking algorithms. However, as shown in the Tables of statistics (Table 5.7 and 5.8), the performance of *k*-core has significantly improved, whereas that of PageRank only slightly improved. This difference can be explained by the fact that *k*-core has been found to be most effective in identifying super-spreaders in previous studies (S. Pei et al., 2014). Furthermore, the Twitter data set has improved more than the Facebook data set because Twitter supports community culture more than Facebook. Twitter members can easily follow their common nodes of interest, whereas Facebook emphasizes mutual friendship rather than unidirectional linking.

DFIS and PSA use two types of association, namely, consistent association and inconsistent association, to find missing values in a soft set. In consistent association, corresponding elements have the same values (0 for 0 and 1 for 1), as shown in Equation (5.2). By contrast, in inconsistent association, the corresponding elements of the columns have compliment values (i.e., 1 for 0 and 0 for 1). For example, column 1 of Table 5.4 exhibits a consistent association with column 6; that is, most of the corresponding values in both columns are the same, whereas the same column 1 has an inconsistent association with column 2. Notably, unlike in consistency, the complement values of the corresponding cells are selected as the predicted values in inconsistency. However, it is suggested that only consistent association is effective for link prediction in OSNs, and inconsistent association should not be applied. Inconsistent association should not be considered because it finds the dissimilarities between prime nodes with respect to followers, whereas we are looking for matches in their followers only. This study uses the term "association" instead of "consistent association" in this article to avoid confusion. Moreover, proposed algorithm does not select any of the inverse or inconsistent association because Equation (5.2) selects only consistent associations. For inconsistent associations, the relations should be changed, and the equal sign (=) should be replaced with the not equal sign (≠). For example, if the relation $CN_{1k} = \{x \mid F(x_1) \neq F(x_k)\}$ is used instead of Equation (5.2), then the algorithm will find k to be column 2, which is inconsistent with column 1 in Table 5.4. By contrast, Equation (5.2) calculates the similarity between linked nodes only (with values equal to 1) because unlinked nodes (with values equal to 0) are already rendered unknown by assigning * to them.

Link prediction by finding the similarity between nodes may be applied if the associated nodes are not actually linked at an instance but are more likely to link to each other. In such cases, a complete network growth through the association between nodes is more reliable in identifying important links and expected influential spreaders.

Two drawbacks of PSA, namely, high computational complexity and rare case false association, have been avoided using the current method. The computational complexity of PSA is $O(n^4)$, and calculating the association between all the nodes in a network is infeasible when handling big data sets such as those used in this study. The Facebook data set has 63,520 nodes connected to 59,222 nodes; thus, a 63,520×59,222 table has to be constructed, and the association for each node with all the other nodes has to be calculated individually. Similarly, the Twitter data set contained 456,626 nodes connected to 370,341 nodes; thus, the BIS of the Twitter data set is approximately 45 times as large as that of the Facebook data set. Small clusters of 2,500 nodes are used for both data sets to avoid this experimental complexity. The average in degree of the Facebook data set is 23.3, and its outdegree is 24.9; thus, the average size of the BIS is 107×100 for each cluster among the 592 clusters. For the Twitter data set, which has an in degree of 32.5 and an out degree of 40.1, the average size of the BIS is 77×62 per cluster out of the 5,942 total clusters. The average size of these clusters is approximately 18 times as large as that of the in/out average degree cluster of the Facebook data set and nearly four times as large as that of the Twitter data set. These calculations suggest that the clusters have sufficient average margins for finding similar nodes, and computational complexity is reduced.

The second PSA drawback of spreading false association and false results can also be avoided using this clustering technique. Although prediction through association is the most favorable method, it is not 100% accurate. In some cases, a real association between the nodes may not actually exist, and the links predicted though association may constitute false links. Such false links can be tolerated within a minimal extent in a large network, but in huge quantities, they are likely to affect network quality and precision. From Equation (5.3), the consistency degree (CD) is the maximum value for the *j*-th column. If the CD value is false for any *j*-th column, particularly for the initial values of *j,* then a false value can be selected by the algorithm for the *k*-th column in any or all of the succeeding iterations, thereby yielding another set of false values based on false values. In the case of a larger BIS, the spreading of these false values are more likely to affect all the predicted values, whereas in small clusters, they will disappear automatically by the end of executing the running cluster with false association. Therefore, a false association affects only a small cluster, and the probability of it spreading is minimized through clustering. Consequently, this clustering technique provides an initial solution for the shortcomings of PSA.

The improvement in the accuracy of the ranking algorithms after link prediction shows that a number of new links are identified during prediction. These links are the present in the diffusion data sets but missing in the extracted network data set. The analysis of both data sets shows that 11,129 new links are predicted in the Facebook network and 445 new links in the Twitter network. Although 445 nodes account for approximately 1/3200 only of the total predicted nodes in the Twitter data set, their role cannot be disregarded. This scenario verifies the approach developed by Li et al. (D. Li et al., 2016), which is integrated into the proposed method.

CD is the ratio between the number of consistent nodes and the total number of known nodes, and it ranges between 0 and 1 (i.e. $1 \geq CD \geq 0$). The higher the value of CD, the more is the similarity between the nodes under consideration, and vice versa. Threshold λ filters similar nodes and can be selected based on filtration size (requirement). Its value ranges between 0 and 1 and represents the similarity (in percent) between two consistent columns. A value closer to 0 indicates less similarities, and vice versa. Threshold selection depends on individual choice, and recommender systems can select a threshold based on their requirements. Recommender systems can also calculate the number of community nodes and recommend prime nodes to the community users based on the threshold value. To select a reliable association, this study recommends a threshold λ value higher than 0.5 to capture associations stronger than 50%.The threshold value is maintained at $\lambda = 0.6$ in the experiments to filter nodes with a minimum of 60% similarity with other nodes in a cluster for both data sets.

5.5 Conclusions and Recommendations

This study discusses the formation of a virtual community and proposes a new identification method for virtual communities in OSNs. Virtual community members are similar in behavior, and this similar behavior is used to solve the link prediction problem. The results are validated by comparing the accuracy rates of the ranking algorithms *k*-core and PageRank through a diffusion graph for two huge real networks, i.e., Facebook and Twitter, before and after the prediction of new links from their corresponding diffusion data sets. The generated results show that the association between prime nodes can be used to solve link prediction problems and explain network growth. The improvement in the accuracy of the ranking algorithms in finding top spreaders validates the proposed method. The division of the BIS into small clusters helps avoid the drawbacks of PSA.

In future studies, a more appropriate and more logical clustering technique can be developed to improve performance results. Furthermore, other prediction features may also be integrated into the proposed method to achieve better performance. Finally, the association between prime nodes in a network can be more accurately determined by considering more than two prime nodes.

CHAPTER 6: CONCLUSION AND FUTURE DIRECTION

6.1 Overview

We examined the tools and techniques used for uncertain data including fuzzy set theory, rough set theory, and soft set theory. Soft set theory is considered the newest and the most efficient tool in handling uncertain data. Soft set theory and its important applications in decision making and parameter reduction were studied in this work. The general causes and effects of incomplete soft sets on their applications were also discussed.

Existing techniques for dealing with incomplete soft sets were reviewed and classified into two categories based on their input for data prediction and recalculation. Approaches that depend on available data for predicting missing values are included in UP category while techniques that depend on equivalency set of aggregates as well as on original available data are included in the PP category. It was shown that PP techniques are unable to recalculate entire values form aggregates in their current form, which called for a novel concept of entire values recalculation from aggregates. The new concept was explained with the help of definitions, mathematical relations, algorithm and a solved example as a proof of concept.

The techniques in the UP category were assessed in terms of their ability to complete the incomplete soft sets, and DFIS was identified as the most suitable technique. DFIS uses association between parameters for data prediction, yet ignores certain association differences. This study has revised the procedure of DFIS operating by proposing an alternative data filling approach PSA that predicts missing values through strongest association first. After implementing DFIS and PSA in MATLAB and predicting deleted values for the benchmark data sets, it was found that the average accuracy of

PSA is higher than that of DFIS. The technique is explained with the help of examples, definitions, algorithm, and mathematical description.

Data prediction through strongest association between parameters in incomplete soft sets was applied to the link prediction problem in online social networks. A new type of network community was detected and named ‘virtual community’ through association between ‘prime nodes’. New links were predicted by implementing the proposed tool in MATLAB for two global OSNs data sets, Facebook and Twitter. The validity of predicted links has been performed by checking the accuracy of ranking algorithms (PageRak and k-Core) for both data sets before and after predicting the new links. The significant improvement in the accuracy of ranking algorithms for completed networks validated the proposed link prediction and shown its efficiency.

6.2 Summary of Results

In relation to recalculating entire missing values from aggregates of the PP category, the concept was illustrated by a simple example of simultaneous and non-simultaneous liner equations and necessary details of definitions and mathematical equations. A proper example of calculating all aggregates and obtaining the original values from these aggregates was solved and explained step by step. The successful re-calculation of entire BIS from the set of aggregates validated the result of this category contribution and explained its procedure.

The second contribution of this study is the prediction of missing values through strongest association in the UP category. The results for this category were obtained by implementing the existing DFIS approach and proposed technique PSA in MATLAB and by testing both for 04 UCI bench mark data sets and LUCAP data set. The significant improvement in the results for each data set validated PSA. The missing values were predicted in a practical example using both approaches. The solved example explained the PSA procedure and showed how the maximum number of values can be predicted through association between parameters instead of probability.

The third contribution of link prediction in online social networks consists of an application of PSA. It was achieved by completing two network data sets of global OSNs Facebook and Twitter though PSA. Facebook data sets have wall posts and Twitter data sets have retweets, from which their actual spreading efficiency was calculated for finding top spreaders. The spreading efficiency of network data sets was also calculated using k-Core and PageRank before and after completion and was then compared with the actual spreading efficiency using the imprecision function. A decrease in the value of the imprecision function after network completion showed higher prediction accuracy.

6.3 Achievement of Objectives

Incomplete soft set handling techniques were discussed and analyzed in this study. Based on their two natural types, all approaches were placed under the PP or UP category. The capability of PP techniques recalculating entire missing values was assessed, and it was shown that these techniques in their current form cannot be used for overall recalculating purposes. A new technique of entire missing values recalculation was introduced in the PP category and a solved example was presented as a proof of concept.

The UP techniques were also analyzed in terms of accuracy, data filling, complexity and integrity. DFIS was identified as the most suitable technique in this category to be used for completing incomplete soft sets. Although DFIS uses association between parameters to predict values and uses probability when association is weak, it treats all associations satisfying the threshold identically. Thus, the role of the strongest association is ignored, which results in low prediction accuracy. After having identified this, the role of the strongest association was considered by introducing the more accurate PSA.

PSA was applied in predicting new links in two OSN data sets (Facebook and Twitter) completed using PSA. In addition to network completion, a new type of network community was found whose nodes have association with each others. The community was named ‘virtual community’ and the associated nodes ‘prime nodes’. Link prediction was validated by finding top spreaders through ranking algorithms k-Core and PageRank before and after network completion. The efficiency of the ranking algorithms was compared with its average spreading efficiency using the related wall posts and retweet data sets. High accuracy in the form of low imprecision function improved the accuracy of ranking algorithms.

6.4 Research Scope and Limitation

The scope of this research incorporates the prediction of missing values and recalculation of entire missing values from aggregates. Prediction is used when there are partial missing values in soft sets while no equivalency set of aggregates is available. In prediction, the partial missing values are found from association between parameters. Missing values are predicted as the corresponding values of its strongly consistent parameters or as the complement of corresponding strongly inconsistent parameters.

If entire values missing and there is no equivalency set of aggregates, such a situation is considered as out of the PSA scope. In such a situation, the availability of a preprocessed equivalency set is mandatory, which enables it to fall within the scope of entire missing values recalculation from aggregates.

Values are recalculated when the equivalency sets are available in the form of diagonals, rows and column aggregates. Partial as well as entire missing values are recalculated from these aggregates using supposition in the Boolean domain.

Two drawbacks of the proposed PSA method can be identified in the form of high computational complexity and rare incorrect values prediction as discussed in this work. Both shortcomings are covered in the application of the proposed approach related to the link prediction problem in OSN by dividing the whole network into small clusters. Using small clusters, the complexity was reduced by calculating inside a small size of BIS instead of the whole network. Incorrect values that were predicted through false association were used inside the cluster only and its effect on other values prediction was avoided using clustering.

6.5 Recommendation and Future Direction

In future, the first proposed method of entire missing value recalculation from aggregates can be used for data compression at the binary level. The second proposed method of data prediction through strongest association is applied in this research to link prediction in OSN, which can be further applied to link and data prediction in other domains of medical and social sciences. The last proposed method of link prediction through association between prime nodes can be implemented in network recommender systems for OSN growth. More accurate results are expected by considering more than two primes nodes for finding association between them.

REFERENCES

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social networks, 25*(3), 211-230.

Agarwal, M., Biswas, K. K., & Hanmandlu, M. (2013). Generalized intuitionistic fuzzy soft sets with applications in decision-making. *Applied soft computing, 13*(8), 3552-3566.

Ahlqvist, T., Bäck, A., Halonen, M., & Heinonen, S. (2008). Social media roadmaps. *Helsinki: Edita Prima Oy*.

Ahmad, B., & Kharal, A. (2009). On fuzzy soft sets. *Advances in Fuzzy Systems, 2009*.

Akdag, M., & Ozkan, A. (2014). On soft β-open sets and soft β-continuous functions. *The Scientific World Journal, 2014*.

Aktaş, H., & Çağman, N. (2007). Soft sets and soft groups. *Information Sciences, 177*(13), 2726-2735.

Alcantud, J. C. R. (2015). Fuzzy soft set based decision making: a novel alternative approach.

Alcantud, J. C. R. (2016). A novel algorithm for fuzzy soft set based decision making from multiobserver input parameter data set. *Information Fusion, 29*, 142-148.

Alhazaymeh, K., & Hassan, N. (2012). Interval-valued vague soft sets and its application. *Advances in Fuzzy Systems, 2012*, 15.

Ali, M. I., Feng, F., Liu, X., Min, W. K., & Shabir, M. (2009). On some new operations in soft set theory. *Computers & Mathematics with Applications, 57*(9), 1547-1553.

Ali, M. I., Mahmood, T., Rehman, M. M. U., & Aslam, M. F. (2015). On lattice ordered soft sets. *Applied soft computing, 36*, 499-505.

Ali, M. I., & Shabir, M. (2014). Logic connectives for soft sets and fuzzy soft sets. *IEEE Transactions on Fuzzy Systems, 22*(6), 1431-1442.

Ali, M. I., Shabir, M., & Naz, M. (2011). Algebraic structures of soft sets associated with new operations. *Computers & Mathematics with Applications, 61*(9), 2647-2654.

Alkhazaleh, S. (2015). The Multi-Interval-Valued Fuzzy Soft Set with Application in Decision Making. *Applied Mathematics, 6*(08), 1250.

Alkhazaleh, S., & Salleh, A. R. (2012). Soft expert sets. *Advances in Decision Sciences, 2011*.

Aslam, M., & Abdullah, S. (2013). Bipolar Fuzzy Soft sets and its applications in decision making problem. *arXiv preprint arXiv:1303.6932*.

Atanassov, K. T. (1986). Intuitionistic fuzzy sets. *Fuzzy Sets and Systems, 20*(1), 87-96.

Aygünoğlu, A., & Aygün, H. (2012). Some notes on soft topological spaces. *Neural computing and Applications, 21*(1), 113-119.

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). *Everyone's an influencer: quantifying influence on twitter.* Paper presented at the Proceedings of the fourth ACM international conference on Web search and data mining.

Basu, T. M., Mahapatra, N. K., & Mondal, S. K. (2012). A balanced solution of a fuzzy soft set based decision making problem in medical science. *Applied soft computing, 12*(10), 3260-3275.

Batagelj, V., & Zaversnik, M. (2003). An O (m) algorithm for cores decomposition of networks. *arXiv preprint cs/0310049*.

Bedi, P., & Sharma, C. (2016). Community detection in social networks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 6*(3), 115-135.

Bellinger, G., Castro, D., & Mills, A. (2004). Data, information, knowledge, and wisdom.

Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks, 56*(18), 3825-3833.

Cagman, N., & Enginoglu, S. (2012). Fuzzy soft matrix theory and its application in decision making. *Iranian Journal of Fuzzy Systems, 9*(1), 109-119.

Çağman, N., & Enginoğlu, S. (2010a). Soft matrix theory and its decision making. *Computers & Mathematics with Applications, 59*(10), 3308-3314.

Çağman, N., & Enginoğlu, S. (2010b). Soft set theory and uni–int decision making. *European Journal of Operational Research, 207*(2), 848-855.

Cagman, N., Enginoglu, S., & Citak, F. (2011). Fuzzy soft set theory and its applications. *Iranian Journal of Fuzzy Systems*.

Cagman, N., Enginoglu, S., & Citak, F. (2011). Fuzzy soft set theory and its applications. *Iranian Journal of Fuzzy Systems, 8*(3), 137-147.

Çağman, N., Karataş, S., & Enginoglu, S. (2011). Soft topology. *Computers & Mathematics with Applications, 62*(1), 351-358.

Çelik, Y., & Yamak, S. (2013). Fuzzy soft set theory applied to medical diagnosis using fuzzy arithmetic operations. *Journal of Inequalities and Applications, 2013*(1), 1-9.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM, 10*(10-17), 30.

Chen, B. (2013). Soft semi-open sets and related properties in soft topological spaces. *Appl. Math. Inf. Sci, 7*(1), 287-294.

Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C., & Zhou, T. (2012). Identifying influential nodes in complex networks. *Physica A: Statistical Mechanics and its Applications, 391*(4), 1777-1787.

Chen, D., Tsang, E., Yeung, D. S., & Wang, X. (2005). The parameterization reduction of soft sets and its applications. *Computers & Mathematics with Applications, 49*(5), 757-763.

Chen, W., Cheng, S., He, X., & Jiang, F. (2012). *Influencerank: An efficient social influence measurement for millions of users in microblog.* Paper presented at the Cloud and Green Computing (CGC), 2012 Second International Conference on.

Danjuma, S., Ismail, M. A., & Herawan, T. (2017). An Alternative Approach to Normal Parameter Reduction Algorithm for Soft Set Theory. *IEEE Access*.

Das, S., & Kar, S. (2014). Group decision making in medical system: An intuitionistic fuzzy soft set approach. *Applied soft computing, 24*, 196-211.

De Domenico, M., Lima, A., Mougel, P., & Musolesi, M. (2013). The anatomy of a scientific rumor. *Scientific reports, 3*.

Deli, I., & Karataş, S. (2016). Interval valued intuitionistic fuzzy parameterized soft set theory and its decision making. *Journal of Intelligent & Fuzzy Systems, 30*(4), 2073-2082.

Dinda, B., Bera, T., & Samanta, T. (2010). Generalised intuitionistic fuzzy soft sets and its application in decision making. *arXiv preprint arXiv:1010.2468*.

Duan, L., Aggarwal, C., Ma, S., Hu, R., & Huai, J. (2016). *Scaling up Link Prediction with Ensembles.* Paper presented at the Proceedings of the Ninth ACM International Conference on Web Search and Data Mining.

Feng, F. (2009). *Generalized rough fuzzy sets based on soft sets.* Paper presented at the Intelligent Systems and Applications, 2009. ISA 2009. International Workshop on.

Feng, F., Jun, Y. B., Liu, X., & Li, L. (2010). An adjustable approach to fuzzy soft set based decision making. *Journal of Computational and Applied Mathematics, 234*(1), 10-20.

Feng, F., Jun, Y. B., & Zhao, X. (2008). Soft semirings. *Computers & Mathematics with Applications, 56*(10), 2621-2628.

Feng, F., Li, C., Davvaz, B., & Ali, M. I. (2010). Soft sets combined with fuzzy sets and rough sets: a tentative approach. *Soft Computing, 14*(9), 899-911.

Feng, F., Li, Y., & Leoreanu-Fotea, V. (2010). Application of level soft sets in decision making based on interval-valued fuzzy soft sets. *Computers & Mathematics with Applications, 60*(6), 1756-1767.

Feng, F., Liu, X., Leoreanu-Fotea, V., & Jun, Y. B. (2011). Soft sets and soft rough sets. *Information Sciences, 181*(6), 1125-1137.

Fortunato, S. (2010). Community detection in graphs. *Physics reports, 486*(3), 75-174.

Gau, W.-L., & Buehrer, D. J. (1993). Vague sets. *IEEE transactions on systems, man, and cybernetics, 23*(2), 610-614.

Güneş, İ., Gündüz-Öğüdücü, Ş., & Çataltepe, Z. (2016). Link prediction using time series of neighborhood-based node similarity scores. *Data Mining and Knowledge Discovery, 30*(1), 147-180.

Herawan, T. (2010). *Soft set-based decision making for patients suspected influenza-like illness.* Paper presented at the International Journal of Modern Physics: Conference Series.

Herawan, T. (2012). *Soft Set-Based Decision Making for Patients Suspected Influenza-Like Illness.* Paper presented at the International Journal of Modern Physics: Conference Series.

Herawan, T., & Deris, M. M. (2009a). *A direct proof of every rough set is a soft set.* Paper presented at the Modelling & Simulation, 2009. AMS'09. Third Asia International Conference on.

Herawan, T., & Deris, M. M. (2009b). On multi-soft sets construction in information systems *Emerging Intelligent Computing Technology and Applications. With Aspects of Artificial Intelligence* (pp. 101-110): Springer.

Herawan, T., & Deris, M. M. (2011). A soft set approach for association rules mining. *Knowledge-Based Systems, 24*(1), 186-195.

Hussain, S., & Ahmad, B. (2011). Some properties of soft topological spaces. *Computers & Mathematics with Applications, 62*(11), 4058-4067.

Isa, A. M., Rose, A. N. M., & Deris, M. M. (2011). Dominance-based soft set approach in decision-making analysis *Advanced Data Mining and Applications* (pp. 299-310): Springer.

Jabeur, L. B., Tamine, L., & Boughanem, M. (2012). *Active microbloggers: identifying influencers, leaders and discussers in microblogging networks.* Paper presented at the International Symposium on String Processing and Information Retrieval.

Jiang, Y., Liu, H., Tang, Y., & Chen, Q. (2011). Semantic decision making using ontology-based soft sets. *Mathematical and Computer Modelling, 53*(5), 1140-1149.

Jiang, Y., Tang, Y., & Chen, Q. (2011). An adjustable approach to intuitionistic fuzzy soft sets based decision making. *Applied Mathematical Modelling, 35*(2), 824-836.

Jiang, Y., Tang, Y., Chen, Q., Liu, H., & Tang, J. (2010). Interval-valued intuitionistic fuzzy soft sets and their properties. *Computers & Mathematics with Applications, 60*(3), 906-918.

Jun, Y. B., Lee, K. J., & Park, C. H. (2009). Soft set theory applied to ideals in d-algebras. *Computers & Mathematics with Applications, 57*(3), 367-378.

Jun, Y. B., & Park, C. H. (2008). Applications of soft sets in ideal theory of BCK/BCI-algebras. *Information Sciences, 178*(11), 2466-2475.

Kahraman, C., Onar, S. C., & Oztaysi, B. (2015). Fuzzy multicriteria decision-making: a literature review. *International Journal of Computational Intelligence Systems, 8*(4), 637-666.

Kalaichelvi, A., & Malini, P. H. (2011a). Application Of Fuzzy Soft Sets To Investment Decision Making Problem. *Int. J. of Mathematical Sciences and Applications, 1*(3).

Kalaichelvi, A., & Malini, P. H. (2011b). Application of fuzzy soft sets to investment decision making problem. *Internal Journal of Mathematical Sciences and Applications, 1*(3), 1583-1586.

Kalayathankal, S. J., & Singh, G. S. (2010). A fuzzy soft flood alarm model. *Mathematics and Computers in Simulation, 80*(5), 887-893.

Kalayathankal, S. J., & Suresh Singh, G. (2010). A fuzzy soft flood alarm model. *Mathematics and Computers in Simulation, 80*(5), 887-893.

Kannan, K. (2012). Soft generalized closed sets in soft topological spaces. *Journal of theoretical and applied information technology, 37*(1), 17-21.

Kim, E. S., & Han, S. S. (2009). *An analytical way to find influencers on social networks and validate their effects in disseminating social games.* Paper presented at the Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature physics, 6*(11), 888-893.

Kong, Z., Gao, L., & Wang, L. (2009). Comment on “A fuzzy soft set theoretic approach to decision making problems”. *Journal of Computational and Applied Mathematics, 223*(2), 540-542.

Kong, Z., Gao, L., Wang, L., & Li, S. (2008). The normal parameter reduction of soft sets and its algorithm. *Computers & Mathematics with Applications, 56*(12), 3029-3037.

Kong, Z., Wang, L., & Wu, Z. (2011). Application of fuzzy soft set in decision making problems based on grey theory. *Journal of Computational and Applied Mathematics, 236*(6), 1521-1530.

Kong, Z., Zhang, G., Wang, L., Wu, Z., Qi, S., & Wang, H. (2014). An efficient decision making approach in incomplete soft set. *Applied Mathematical Modelling, 38*(7), 2141-2150.

Kossinets, G. (2006). Effects of missing data in social networks. *Social networks, 28*(3), 247-268.

Lauw, H., Shafer, J. C., Agrawal, R., & Ntoulas, A. (2010). Homophily in the digital world: A LiveJournal case study. *IEEE Internet Computing, 14*(2), 15-23.

Li, D., Zhang, Y., Xu, Z., Chu, D., & Li, S. (2016). Exploiting Information Diffusion Feature for Link Prediction in Sina Weibo. *Scientific reports, 6*.

Li, Z., Wen, G., & Xie, N. (2015). An approach to fuzzy soft sets in decision making based on grey relational analysis and Dempster–Shafer theory of evidence: An application in medical diagnosis. *Artificial Intelligence in Medicine*.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology, 58*(7), 1019-1031.

Lin, T. (1998). Granular computing on binary relations II: Rough set representations and belief functions. *Rough Sets In Knowledge Discovery, 1*, 121-140.

Liu, Y., Tang, M., Zhou, T., & Do, Y. (2015). Core-like groups result in invalidation of identifying super-spreader by k-shell decomposition. *Scientific reports, 5*.

Lü, L., Zhang, Y.-C., Yeung, C. H., & Zhou, T. (2011). Leaders in social networks, the delicious case. *PloS one, 6*(6), e21202.

Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications, 390*(6), 1150-1170.

Ma, X., Sulaiman, N., Qin, H., Herawan, T., & Zain, J. M. (2011). A new efficient normal parameter reduction algorithm of soft sets. *Computers & Mathematics with Applications, 62*(2), 588-598.

Mahanta, J., & Das, P. (2017). Fuzzy soft topological spaces. *Journal of Intelligent & Fuzzy Systems, 32*(1), 443-450.

Maji, P., Biswas, R., & Roy, A. (2003). Soft set theory. *Computers & Mathematics with Applications, 45*(4-5), 555-562.

Maji, P., Roy, A. R., & Biswas, R. (2002). An application of soft sets in a decision making problem. *Computers & Mathematics with Applications, 44*(8), 1077-1083.

Maji, P. K. (2009). *More on intuitionistic fuzzy soft sets.* Paper presented at the International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing.

Maji, P. K. (2012). A neutrosophic soft set approach to a decision making problem. *Annals of Fuzzy Mathematics and Informatics*, *3*(2), 313-319.

Maji, P. K. (2013). Neutrosophic soft set. *Annals of Fuzzy Mathematics and Informatics*, *5*(1), 157-168.

Maji, P. K., BISWAS, R., & Roy, A. (2001). Fuzzy soft sets.

Majumdar, P., & Samanta, S. K. (2010a). Generalised fuzzy soft sets. *Computers & Mathematics with Applications, 59*(4), 1425-1432.

Majumdar, P., & Samanta, S. K. (2010b). On soft mappings. *Computers & Mathematics with Applications, 60*(9), 2666-2672.

Mamat, R., Herawan, T., & Deris, M. M. (2013). MAR: Maximum Attribute Relative of soft set for clustering attribute selection. *Knowledge-Based Systems*, *52*, 11-20.

Min, B., Liljeros, F., & Makse, H. A. (2015). Finding influential spreaders from human activity beyond network location. *PloS one, 10*(8), e0136831.

Min, W. K. (2011). A note on soft topological spaces. *Computers & Mathematics with Applications*, *62*(9), 3524-3528.

Mohd Rose, A. N., Hassan, H., Awang, M. I., Mahiddin, N. A., Mohd Amin, H., & Deris, M. M. (2011). Solving incomplete datasets in soft set using supported sets and aggregate values. *Procedia Computer Science*, *5*, 354-361.

Molodtsov, D. (1999). Soft set theory—first results. *Computers & Mathematics with Applications*, *37*(4), 19-31.

Moore, R., & Lodwick, W. (2003). Interval analysis and fuzzy set theory. *Fuzzy Sets and Systems*, *135*(1), 5-9.

Morone, F., & Makse, H. A. (2015). Influence maximization in complex networks through optimal percolation. *Nature*.

Nazmul, S., & Samanta, S. (2012). Neighbourhood properties of soft topological spaces. *Annals of Fuzzy Mathematics and Informatics, 6*, 1-15.

Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical review E, 64*(2), 025102.

Palla, G., Barabási, A.-L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature, 446*(7136), 664-667.

Pawlak, Z. (1982). Rough sets. *International Journal of Computer & Information Sciences, 11*(5), 341-356.

Pawlak, Z. (1982). Rough Sets. *International Journal of Injonation and Computer Sciences, 11*, 341-356.

Pawlak, Z. (1998). Rough set theory and its applications to data analysis. *Cybernetics & Systems, 29*(7), 661-688.

Pawlak, Z. (2012). *Rough Sets: Theoretical Aspects of Reasoning about Data*: Springer Netherlands.

Pei, D., & Miao, D. (2005). *From soft sets to information systems.* Paper presented at the Granular Computing, 2005 IEEE International Conference on.

Pei, S., Muchnik, L., Andrade Jr, J. S., Zheng, Z., & Makse, H. A. (2014). Searching for superspreaders of information in real-world social media. *Scientific reports, 4*.

Pei, S., Muchnik, L., Tang, S., Zheng, Z., & Makse, H. A. (2015). Exploring the complex pattern of information spreading in online blog communities. *PloS one, 10*(5), e0126894.

Peng, C., Kolda, T. G., & Pinar, A. (2014). Accelerating community detection by using k-core subgraphs. *arXiv preprint arXiv:1403.2226*.

Polat, N. C., & Tanay, B. (2016). A Method for Decision Making Problems by Using Graph Representation of Soft Set Relations. *ISTANBUL COMMERCE UNIVERSITY*, 181.

Qin, H., Ma, X., Herawan, T., & Zain, J. M. DFIS: A novel data filling approach for an incomplete soft set. *International Journal of Applied Mathematics and Computer Science, 22*(4), 817-828.

Qin, H., Ma, X., Herawan, T., & Zain, J. M. (2011a). An adjustable approach to interval-valued intuitionistic fuzzy soft sets based decision making *Intelligent Information and Database Systems* (pp. 80-89): Springer.

Qin, H., Ma, X., Herawan, T., & Zain, J. M. (2011b). Data filling approach of soft sets under incomplete information *Intelligent Information and Database Systems* (pp. 302-311): Springer.

Qin, H., Ma, X., Herawan, T., & Zain, J. M. (2012a). DFIS: A novel data filling approach for an incomplete soft set. *Int. J. Appl. Math. Comput. Sci, 22*(4), 817-828.

Qin, H., Ma, X., Herawan, T., & Zain, J. M. (2012b). DFIS: a novel data filling approach for an incomplete soft set. *International Journal of Applied Mathematics and Computer Science, 22*(4), 817-828.

Qin, H., Ma, X., Zain, J. M., & Herawan, T. (2012). A novel soft set approach in selecting clustering attribute. *Knowledge-Based Systems, 36*, 139-145.

Qin, H. W., Ma, X. Q., Herawan, T., & Zain, J. M. (2012). Dfis: A Novel Data Filling Approach for an Incomplete Soft Set. *International Journal of Applied Mathematics and Computer Science, 22*(4), 817-828. doi: 10.2478/v10006-012-0060-3

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America, 101*(9), 2658-2663.

Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011). *Influence and passivity in social media.* Paper presented at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases.

Rose, A. N. M., Hassan, H., Awang, M. I., Herawan, T., & Deris, M. M. (2011). Solving Incomplete Datasets in Soft Set Using Parity Bits of Supported Sets *Ubiquitous Computing and Multimedia Applications* (pp. 33-43): Springer.

Roy, A. R., & Maji, P. (2007). A fuzzy soft set theoretic approach to decision making problems. *Journal of Computational and Applied Mathematics, 203*(2), 412-418.

Sezgin, A., & Atagün, A. O. (2011). Soft groups and normalistic soft groups. *Computers & Mathematics with Applications, 62*(2), 685-698.

Shabir, M., & Naz, M. (2011). On soft topological spaces. *Computers & Mathematics with Applications, 61*(7), 1786-1799.

Shao, Y., & Qin, K. (2012). Fuzzy soft sets and fuzzy soft lattices. *International Journal of Computational Intelligence Systems, 5*(6), 1135-1147.

Silva, A., Guimarães, S., Meira Jr, W., & Zaki, M. (2013). *ProfileRank: finding relevant content and influential users based on information diffusion.* Paper presented at the Proceedings of the 7th Workshop on Social Network Mining and Analysis.

Sulaiman, N. H., & Mohamad, D. (2013). Multiaspect soft sets. *Advances in Fuzzy Systems, 2013*, 1.

Sun, P. G. (2016). Imbalance problem in community detection. *Physica A: Statistical Mechanics and its Applications, 457*, 364-376.

Sutoyo, E., Mungad, M., Hamid, S., & Herawan, T. (2016). An Efficient Soft Set-Based Approach for Conflict Analysis. *PloS one, 11*(2), e0148837.

Tanay, B., & Kandemir, M. B. (2011). Topological structure of fuzzy soft sets. *Computers & Mathematics with Applications, 61*(10), 2952-2957.

Tripathy, B., Mohanty, R., & Sooraj, T. (2016). *On intuitionistic fuzzy soft set and its application in group decision making.* Paper presented at the Emerging Trends in Engineering, Technology and Science (ICETETS), International Conference on.

Wang, F., Li, X., & Chen, X. (2014). Hesitant fuzzy soft set and its applications in multicriteria decision making. *Journal of Applied Mathematics, 2014*.

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). *Twitterrank: finding topic-sensitive influential twitterers.* Paper presented at the Proceedings of the third ACM international conference on Web search and data mining.

Xiao, Z., Gong, K., & Zou, Y. (2009). A combined forecasting approach based on fuzzy soft sets. *Journal of Computational and Applied Mathematics, 228*(1), 326-333.

Xu, W., Ma, J., Wang, S., & Hao, G. (2010). Vague soft sets and their properties. *Computers & Mathematics with Applications, 59*(2), 787-794.

Yang, X., Yu, D., Yang, J., & Wu, C. (2007). Generalization of soft set theory: from crisp to fuzzy case *Fuzzy Information and Engineering* (pp. 345-354): Springer.

Yang, Y., Tan, X., & Meng, C. (2013). The multi-fuzzy soft set and its application in decision making. *Applied Mathematical Modelling, 37*(7), 4915-4923.

Yao, Y. (1998). Relational interpretations of neighborhood operators and rough set approximation operators. *Information Sciences, 111*(1), 239-259.

Yuksel, S., Dizman, T., Yildizdan, G., & Sert, U. (2013). Application of soft sets to diagnose the prostate cancer risk. *Journal of Inequalities and Applications, 2013*(1), 1-11.

Zadeh, L. A. (1965). Fuzzy Set. *Information and Controle, 8*.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control, 8*(3), 338-353.

Zhan, W., Guan, J., Chen, H., Niu, J., & Jin, G. (2016). Identifying overlapping communities in networks using evolutionary method. *Physica A: Statistical Mechanics and its Applications, 442*, 182-192.

Zhang, X. (2014). On interval soft sets with applications. *International Journal of Computational Intelligence Systems, 7*(1), 186-196.

Zhang, Z. (2012). A rough set approach to intuitionistic fuzzy soft set based decision making. *Applied Mathematical Modelling, 36*(10), 4605-4633.

Zimmerman, H. (1991). Fuzzy Set Theory and Its Applications.

Zimmermann, H.-J. (2001). *Fuzzy set theory—and its applications*: Springer Science & Business Media.

Zimmermann, H.-J. (2014). Fuzzy Set Theory-and Its Applications.

Zimmermann, H. (1991). Fuzzy set theory: and its applications.

Zorlutuna, I., Akdag, M., Min, W., & Atmaca, S. (2012). Remarks on soft topological spaces. *Annals of Fuzzy Mathematics and Informatics, 3*(2), 171-185.

Zou, Y., & Xiao, Z. (2008). Data analysis approaches of soft sets under incomplete information. *Knowledge-Based Systems, 21*(8), 941-945.

LIST OF PUBLICATIONS

1. Khan, M.S., Wahab, A. W.A., Herawan, T., Mujtaba, G., Danjuma, S., & Al-Garadi, M. A. (2016). **Virtual community detection through the association between prime nodes in online social networks and its application to ranking algorithms.** *IEEE Access*, *4*, 9614-9624.

2. Khan, M. S., Herawan, T., Wahab, A. W. A., Mujtaba, G., & Al-Garadi, M. A. (2017). **Concept of entire Boolean values recalculation from aggregates in the preprocessed category of incomplete soft sets.** *IEEE Access*, *5*, 11444-11454.

3. Khan, M. S., Al-Garadi, M. A., Wahab, A. W. A., & Herawan, T. (2016). **An alternative data filling approach for prediction of missing data in soft sets (ADFIS).** *SpringerPlus*, *5*(1), 1348.