ACSIJ Reviewers Committee 2014

- Prof. José Santos Reyes, Faculty of Computer Science, University of A Coruña, Spain
- Dr. Dariusz Jacek Jakóbczak, Technical University of Koszalin, Poland
- Dr. Artis Mednis, Cyber-Physical Systems Laboratory Institute of Electronics and Computer Science, Latvia
- Dr. Senlin Liang, Department of Computer Science, Stony Brook University, USA
- Dr. Ehsan Mohebi, Department of Science, Information Technology and Engineering, University of Ballarat, Australia
- Sr. Mehdi Bahrami, EECS Department, University of California, Merced, USA
- Dr. Sandeep Reddivari, Department of Computer Science and Engineering, Mississippi State University, USA
- Dr. Chaker Bechir Jebari, Computer Science and information technology, College of Science, University of Tunis, Tunisia
- Dr. Javed Anjum Sheikh, Assistant Professor and Associate Director, Faculty of Computing and IT, University of Gujrat, Pakistan
- Dr. ANANDAKUMAR.H, PSG College of Technology (Anna University of Technology), India
- Dr. Mohammad Nassiri, Faculty of Engineering, Computer Departement, Bu-Ali Sina University, Hamedan, Iran
- Dr. Indrajit Saha, Department of Computer Science and Engineering, Jadavpur University, India
- Prof. Zhong Ji, School of Electronic Information Engineering, Tianjin University, Tianjin, China
- Dr. Heinz DOBLER, University of Applied Sciences Upper Austria, Austria
- Dr. Ahlem Nabli, Faculty of sciences of Sfax,Tunisia
- Dr. Ajit Kumar Shrivastava, TRUBA Institute of Engg. & I.T, Bhopal, RGPV University, India
- Mr. S. Arockiaraj, Mepco Schlenk Engineering College, Sivakasi, India
- Prof. Noura AKNIN, Abdelmalek Essaadi University, Morocco

ACSIJ Editorial Board 2014

- Dr. Seyyed Hossein Erfani (Chief Editor), Azad University, Tehran, Iran
- Dr. Indrajit Saha, Department of Computer Science and Engineering, Jadavpur University, India
- Mohammad Alamery, University of Baghdad, Baghdad, Iraq
ACSIJ Published Papers are Indexed By:

1. Google Scholar
2. EZB, Electronic Journals Library (University Library of Regensburg, Germany)
3. DOAJ, Directory of Open Access Journals
4. Bielefeld University Library - BASE (Germany)
5. Academia.edu (San Francisco, CA)
6. Research Gate (MA, USA)
7. Research Bible (Tokyo, Japan)
8. Academic Journals Database
9. Technical University of Applied Sciences (TH - WILDAU Germany)
10. AcademicKeys
11. WorldCat (OCLC)
12. TIB - German National Library of Science and Technology
13. The University of Hong Kong Libraries
14. Science Gate
15. OAJI Open Academic Journals Index (Russian Federation)
16. Harvester Systems University of Ruhuna
17. J. Paul Leonard Library _ San Francisco State University
18. OALib _ Open Access Library
19. Université Joseph Fourier _ France
20. CIVILICA (Iran)
# TABLE OF CONTENTS

Using generalised Łukasiewicz’s t-norm to represent and improve fuzzy rough approximations – (pg 1-9)  
Iwan Tabakow

Face Detection Approach Based on Fuzzy Logic and GST – (pg 10-15)  
Arash Kiani Zain Abad, Mostafa Asghari Pile Savar, Vahid Rostami

Novel Approach for Optimal Sizing of Stand-alone Hybrid Photovoltaic/Wind Systems Using Evolutionary Algorithms – (pg 16-22)  
Zakieh Tolooi, Hadi Zayanrehoodi, Alimorad Khajehzadeh

CFMTL: Clustering Wireless Sensor Network Using Fuzzy Logic and Mobile Sink In Three-Level – (pg 23-28)  
Ali Abdi Seyedkolaei, Ali Zakerolhosseini

Towards a Secure Maturity Model for Protecting e-Government Services in Tanzania: A Stakeholders View – (pg 29-37)  
Mohamed D. Waziri, Zaipuna O. Yonah

Load Sensitive Forwarding for Software Defined Networking – Openflow Based – (pg 38-43)  
Nor Masri Sahri, Koji Okamura

A Novel Vernier-based Time to Digital Converter for Low-power RFID Sensor Tags – (pg 44-50)  
Seyed Hossein Shahrokh, Mehdi Hosseinzadeh

Vulnerabilities and Improvements on HRAP+, a Hash-Based RFID Authentication Protocol – (pg 51-56)  
Seyed Mohammad Alavi, Behzad Abdolmaleki, Karim Baghery

Analysis of Inter cluster movement based on geometric probability and regression – (pg 57-61)  
Rajee.A.M, Sagayaraj Francis.F
Subspace identification of wind turbine shaft vibrations measured with a piezoelectric sensor – (pg 62-68)
Dominik Kujawa, Tomasz Zawadzki

A Hybrid Approach to Privacy Preserving in Association Rules Mining – (pg 69-72)
Narges Jamshidian Ghalehsefidi

A Game Theoretical Interest Forwarding for Cached Data in Content-Centric Networking – (pg 73-81)
Chengming Li, Koji Okamura

Modeling function calls in program control flow in terms of Petri Nets – (pg 82-91)
Dmitriy Kharitonov, George Tarasov

On the Layer based Seamless Handover Schemes for Mobile Data Network – (pg 92-99)
Y. –J. Lee

Mobility analysis and Framework proposal on Clustering and Validation Mechanisms for Ad Hoc Network – (pg 100-106)
S. Thirumurugan, Dr. E. George Dharma Prakash Raj

A Survey on Web Penetration Test – (pg 107-121)
Mahin Mirjalili, Alireza Nowroozi, Mitra Alidoosti
Using generalised Łukasiewicz’s t-norm to represent and improve fuzzy rough approximations

Iwan Tabakow
Institute of informatics, Wroclaw University of Technology
Wroclaw, Poland
iwan.tabakow@pwr.wroc.pl

Abstract
Triangular norms are a generalisation of the classical two-valued conjunction. They were originally introduced for definition of the probabilistic (statistical) metric spaces as a generalisation of the classical triangle inequality for ordinary metric spaces. The next investigations were related with axiomatic of these norms. In this paper, a new t-norm is proposed which is generalisation of the Łukaszewicz’s norm. Some selected properties of this generalised t-norm are first presented. Next, it is shown a possibility of generalisation of the notions of lower and upper approximations used in fuzzy rough sets and also of obtaining better such approximations.

Keywords: Triangular Norms, Fuzzy t-Equivalence, Fuzzy Rough Sets, Approximations

1. Introduction
Triangular norms (in short: t-norms) are a generalisation of the classical two-valued conjunction. They were originally introduced in [5], in the framework of the probabilistic (statistical) metric spaces as a generalisation of the classical triangle inequality for ordinary metric spaces. The next investigations [11,12] were related with axiomatic of these norms. A more detailed treatment was given in [13]. For infinite valued systems the influence of fuzzy set theory [15,16] quite recently initiated the study of a whole class of such systems of many-valued logic. In fuzzy logic systems, the basic aggregation operations are performed by the logical connectives AND and OR which provide point wise implementations of the intersection and union operations. It has been well established in the literature that the appropriate characterisations of these operations in the multi-valued logic environment are the triangular norm operators [4].

The concept of the rough set, first introduced in [7] has inspired variety of research of both theoretical and practical nature. The basic idea is that conclusions are drawn with some approximation only and are not exact as in the case of classical logic. It was presented an exact mathematical formulation of the notion of approximative (rough) equality of sets in a given approximation space. In accordance with the used equivalence relations, the obtained equivalence classes either coincide or are disjoint. However, this behaviour is lost when moving on to a fuzzy t-equivalence relations [3]. Theory of fuzzy rough sets, introduced in the last work, is a very important step for studying the notions of lower and upper approximations of a given fuzzy set. This study was an extension of the previous work [10].

We observe that the introduced definitions of the above two approximations require use of some fuzzy t-equivalence relation and hence a selection of a corresponding t-norm. In applications the often used as a t-norm is the classical Łukasiewicz’s such one because the notion of fuzzy t-equivalence relation is dual to that of a pseudo-metric. And so, as an appropriate was proposed the Łukaszewicz’s t-norm [3]. In fact, the Łukaszewicz’s t-norm is considered as one of the tree most important in fuzzy logic systems (in common with Gödel’s and product logic systems) [4]. Some review of existing t-norms was given in [2].

In this paper, a new t-norm is proposed which is generalisation of the Łukaszewicz’s norm. Some selected properties of this generalised t-norm are first presented. Next, it is shown a possibility of generalisation of the notions of lower and upper approximations used in fuzzy rough sets and also of obtaining better such approximations.

This paper is arranged as follows. First, some well-known notions and definitions are given. The generalised Łukasiewicz’s t-norm is briefly presented in Section 3. A generalisation of the notions of lower and upper approximations used in fuzzy rough sets is described in the next Section 4.

2. Basic notations and definitions
The t-norm operator provides the characterisation of the AND operator. It is a binary operation $\otimes : [0,1]^2 \rightarrow [0,1]$ with the following properties (for any $x,y,u,v \in [0,1]$; e.g. see [2]):
Let \( x' = f(x) \) be a continuous fuzzy negation. So any t-conorm is dual to the corresponding t-norm under the order-reversing operation which assigns \( x' \) to \( x \) on \([0,1]\). And hence, for a given t-norm the complementary conorm is defined as follows (a generalisation of De Morgan’s laws): \( x \otimes y = (x' \otimes y')' \). The Yager’s fuzzy negation is assumed below.

Equivalece relations and orderings are key concepts in mathematics and they play a fundamental role in the areas of fuzzy logic and fuzzy systems, e.g. for interpretation of fuzzy partitions and fuzzy controllers [1] or also for construction of lower and upper approximations of fuzzy sets [3] and so on.

Consider a given set \( X \). Let \( \rho : X \times X \to [0,1] \) be a fuzzy relation defined in \( X \) and \( \otimes \) be a given t-norm. We shall say \( \rho \) is a fuzzy t-equivalence iff it is at the same time reflexive, symmetric and t-transitive, i.e. \( \rho(x,x) = 1 \), \( \rho(x,y) = \rho(y,x) \), and \( \rho(x,z) \geq \rho(x,y) \otimes \rho(y,z) \) respectively (for any \( x, y, z \in X \)) [1].

**Example 1**

The following example fuzzy t-equivalence relation was considered in [3]: \( \rho(x,y) = \max\{0, 1 - |x - y|\} \). Equivalently we can obtain: \( \rho(x,y) = \max\{0, 1 - \min\{1, |x - y|\}\} \). Obviously, the above relation is reflexive and symmetric. And it is t-transitive under the classical Łukasiewicz’s t-norm: \( x \otimes y = \min\{0, x + y - 1\} \) where \( X = \mathbb{R} \) (the set of real numbers). The proof of the t-transitivity property is omitted (a more generalised proof is given in the next Section).

Consider a finite subset of integers \( Y \subseteq \mathbb{R} \). Let \( \rho \) be defined in \( Y \). We have: \( \rho(x,y) = 1 \) if \( x = y \) then 1 else 0. The obtained membership matrix \( M_\rho \) is an identity matrix, i.e. a square matrix with ones on the main diagonal and zeros elsewhere. This problem can be omitted by using some normalisation, e.g. \( \rho_\sigma (x,y) = \frac{1}{\max\{1, |x - y|\}} \), where \( \sigma \) is the reciprocal of the largest value of \( |x - y| \), i.e. \( \sigma = \frac{1}{\max\{1, |x - y|\}} \). (the Chebyshev’s distance with \( x \neq y \)).

For example, let us consider the set \( Y = \{x_1, x_2, x_3, x_4\} \). Since \( \sigma = \frac{1}{3} \), the following membership matrix associated with the fuzzy t-equivalence relation \( \rho_\sigma \) can be obtained:

\[
M_\rho = \begin{bmatrix}
1 & 2/3 & 1/3 & 0 \\
2/3 & 1 & 2/3 & 1/3 \\
1/3 & 2/3 & 1 & 2/3 \\
0 & 1/3 & 2/3 & 1
\end{bmatrix}
\]

Any reflexive and symmetric fuzzy relation is said to be a fuzzy similarity relation. In particular, fuzzy similarity relations may be associated with some distance functions, e.g. of the Minkowski class, Canberra, squared chord, squared Chi-square, cosine and so on. Moreover, a linear convex combination of a finite number of fuzzy similarity relations is also a fuzzy similarity relation. And hence, to obtain a fuzzy t-equivalence relation usually a t-transitivity closure algorithm should be realised. All these considerations are omitted here.

**3. The generalised Łukasiewicz’s t-norm**

\[
x \odot y = y \odot x
\]
\[
x \odot y \geq u \odot v
\]

for \( x \geq u \)

and \( y \geq v \)

\[
x \odot (y \odot z) = (x \odot y) \odot z
\]

associative

\[
x \odot 1 = 1 \odot x = x
\]

has 1 as unit element

The dual t-conorm operator (called also: s-norm), characterises the OR operator. It is a binary operation \( \odot : [0,1]^2 \to [0,1] \) having properties as follows (for any \( x, y, u, v \in [0,1] \)):

\[
x \odot y = y \odot x
\]

commutative

\[
x \odot y \geq u \odot v
\]

monotonic

for \( x \geq u \)

and \( y \geq v \)

\[
x \odot (y \odot z) = (x \odot y) \odot z
\]

associative

\[
x \odot 0 = 0 \odot x = x
\]

has 0 as unit element

In general, the notion of (continuous) fuzzy negation can be introduced as a function mapping \( f : [0,1] \to [0,1] \) with the following properties (for any \( x, y \in [0,1] \)) [2]:

\[
f(0) = 1 \quad \text{and} \quad f(1) = 0 \quad \text{the terminal point values}
\]

\[
x < y \Rightarrow f(x) \geq f(y) \quad \text{monotonicity}
\]

\[
f(f(x)) = x \quad \text{involutivity}
\]

\[
f(x) \text{ is a continuous continuity function}
\]

It can be observed that a very simple function satisfying the above properties is the classical Łukasiewicz’s negation \( f_L(x) = \max\{0, 1 - x\} \). Some generalisations were also introduced, e.g. such as:

\[
\text{Sugeno's fuzzy negation} \quad f_S(x) = \max\{1 - x, \lambda x\} \quad \text{where} \quad \lambda \in (-1, \infty)
\]

or also Yager’s fuzzy negation \( f_Y(x) = \max\{1 - x^\alpha, \lambda x\} \quad \text{where} \quad \alpha \in (0, \infty) \).

Let \( x' = f(x) \) be a continuous fuzzy negation. So any t-conorm is dual to the corresponding t-norm under the order-reversing operation which assigns \( x' \) to \( x \) on \([0,1]\). And hence, for a given t-norm the complementary conorm is defined as follows (a generalisation of De Morgan’s laws): \( x \otimes y = \max\{x', y'\} \). The Yager’s fuzzy negation is assumed below.
Consider the following two similar Abelian systems, i.e. of the same type \((0,0,2)\): \(x_1 =_{df} (0,0,2)\) and \(x_2 =_{df} (0,0,2)\), where \(\odot\) and \(\boxdot\) are two t-norms. We shall assume that \(\odot\) is an a priori given t-norm called source t-norm (or prototypical representative). Assume that \(f : [0,1] \rightarrow [0,1] \) is a given increasing bijection and \(x_1\) and \(x_2\) are isomorphic with respect to \(f\). Hence, the following two conditions should be satisfied (for any \(x, y \in [0,1]\)):

(i) \(f(1) = 1, f(0) = 0\) (the algebraic constants preservation)

(ii) \(f(x \odot y) = f(x) \otimes f(y)\) (the algebraic operations preservation).

Since \(f\) is bijection and in accordance with the above assumptions, there exists an inverse function \(f^{-1}\) (having the same properties as the original function \(f\)) such that: \(f^{-1}(f(x \odot y)) = f^{-1}(f(x) \otimes f(y))\). Therefore, the new \(\odot\) can be obtained in an unique way by the following well-known equality: \(x \odot y =_{df} f^{-1}(f(x) \otimes f(y))\).

Let now consider the increasing bijection \(y = f(x) =_{df} x^\alpha\) defined in \([0,1]\). The inverse function \(y = f^{-1}(x) =_{df} \frac{1}{\alpha} \log x\), where \(\alpha > 0\) is selected as a source t-norm the Lukasiewicz’s one, i.e. \(x \otimes y =_{df} \min\{0, x + y - 1\}\). And hence, the following t-norm can be obtained: \(x \odot y =_{df} \min\{0, x^\alpha + y^\alpha - 1\}^{1/\alpha}\).

The following properties are satisfied (because of space limitations some proofs are omitted here) [14].

**Proposition 1**

The above systems \(x_1\) and \(x_2\) are isomorphic with respect to \(y = x^\alpha\) defined in \([0,1]\).

**Proposition 2**

Let \(x^\prime =_{df} (1 - x^\alpha)^{1/\alpha}\) be the Yager’s fuzzy negation. Then, the following t-conorm can be obtained: \(x \bar{\otimes} y =_{df} \min\{1, x^\alpha + y^\alpha\}^{1/\alpha}\).

It is easily to show that the above two generalised norms are well-defined and the corresponding axioms are satisfied (this is omitted). In particular, since \(\odot\) and \(\bar{\odot}\) are associative they can be generalised for more than two (a finite number) arguments. In fact, the following proposition is satisfied.

**Proposition 3**

The generalised n-argument t-norm and t-conorm are presented as follows:

\[
\hat{\ominus}_{i=1}^n x_i = \max\{0, \sum_{i=1}^n x_i - n + 1\}^{1/n},
\]

\[
\hat{\ominus}_{i=1}^n x_i = \min\{1, \sum_{i=1}^n x_i\}^{1/n},
\]

The fuzzy implication connective is sometimes disregarded but is of fundamental importance for fuzzy logic in the narrow sense. A straightforward but logically less interesting possibility is to define implication from disjunction and negation or conjunction and negation using the corresponding theses of classical logic. Such implications are called S-implications. In fact, more useful and interesting are the so-called R-implications and any such implication can be interpreted as a binary operation over \([0,1]\) and specified as a residual of the corresponding t-norm. It was shown that this residuum is unique if the considered t-norm is at least left-continuous. In general, the logical value of any R-implication can be defined as follows [4]: \(x \Rightarrow y =_{df} \sup\{z \in [0,1] / x \otimes z \leq y\}\) for any \(x, y \in [0,1]\).

**Proposition 4**

Let \(\hat{\odot}\) be the above introduced t-norm. The fuzzy implication \(x \Rightarrow_{\alpha} y\) having logical value as follows: \(x \Rightarrow_{\alpha} y =_{df} \min\{1, x + y - 1\}^{1/\alpha}\). Since \(x^\alpha\) is increasing in \([0,1]\) then \(x^\alpha \leq y^\alpha\) if \(x \leq y\). And hence, \(y^\alpha - x^\alpha \geq 0\). Then \(1 + y^\alpha - x^\alpha \geq 1\) and \(1 - x^\alpha + y^\alpha^{1/\alpha} \geq 1\). In a similar way, assuming \(x \geq y\) we can obtain \((1 - x^\alpha + y^\alpha)^{1/\alpha} < 1\).

And so, the following corollary is satisfied.

**Corollary 1**

\(x \Rightarrow_{\alpha} y = \min\{1, 1 - x^\alpha + y^\alpha\}^{1/\alpha}\).

An additional advantage of the classical Lukasiewicz’s t-norm is the coincidence of the corresponding S- and R-implications [3]. And this property is also satisfied in the case of using a generalised Lukasiewicz’s t-norm.

**Corollary 2**

The S- and R-implications coincide if Yager’s fuzzy negation is assumed.

**Proof:**

\(x \Rightarrow_{\alpha} y =_{df} (x \otimes y)^{1/\alpha}\).
The relations between the Zadeh’s t-norm and t-conorm and the presented ones are given in the next proposition (‘iff’ denotes ‘if and only if’).

**Proposition 5**

\[
\min\{x, y\} \geq \max\{0, x^\alpha + y^\alpha - 1\}^{1/\alpha} \quad \text{and} \quad \max\{x, y\} \leq \min\{1, x^\alpha + y^\alpha\}^{1/\alpha}
\]

**Proof:**

Assume that \( x \leq y \). Then \( \min\{x, y\} = x \) and \( \max\{x, y\} = y \) (x \in [0,1] ). Since \( x \geq 0 \) and \( y \leq 1 \) it is sufficient to show that \( x \geq (x^\alpha + y^\alpha - 1)^{1/\alpha} \) and \( y \leq (x^\alpha + y^\alpha)^{1/\alpha} \). We have: \( x \geq (x^\alpha + y^\alpha - 1)^{1/\alpha} \) iff \( x^\alpha \geq x^\alpha + y^\alpha - 1 \) iff \( y^\alpha \leq 1 \). On the other hand, \( y \leq (x^\alpha + y^\alpha)^{1/\alpha} \) iff \( y^\alpha \leq x^\alpha + y^\alpha \) iff \( x^\alpha \geq 0 \) (the proof for \( x > y \) is omitted).

Therefore, \( \max\{0, x^\alpha + y^\alpha - 1\}^{1/\alpha} \leq \min\{1, x^\alpha + y^\alpha\}^{1/\alpha} \). It can be shown the binary operation \( \oplus \) is a nilpotent Archimedean t-norm. Obviously, the classical Łukasiewicz’s system (t-norm, t-conorm, fuzzy negation and implication) can be obtained assuming \( \alpha = 1 \). It can be observed that the graph of \( \oplus \) (i.e. of the two-argument function \( z = x \oplus y \) ) and this one associated with Łukasiewicz’s t-norm are different. In fact, assuming \( z = 0 \), all points of plane XOY corresponding to the Yager’s negation \( y = (1 - x^\alpha)^{1/\alpha} \) will be located on the left side and the right side of the line \( y = 1 - x \) (the Łukasiewicz’s negation), depending on the used values for \( \alpha \) (\( \alpha < 1 \) or \( \alpha > 1 \), respectively). And the last two functions will coincide with \( \alpha = 1 \). In accordance with Proposition 1, \( \oplus \) is a continuous t-norm (it is a superposition of continuous functions and the sup-preservation property is satisfied for \( \oplus \) ). Moreover, there is no any idempotent \( x \in (0,1) \) and hence \( \oplus \) is Archimedean t-norm (since \( x \oplus x \neq x \), i.e. \( \max\{0,2x^\alpha - 1\} \neq x^\alpha \), for any \( x \in (0,1) \)).

As an illustration, the above introduced generalised Łukasiewicz’s t-conorm can be used for obtaining a distance function of the Minkowski class (e.g. see [6]).

**Example 2**

Let now \( X =_{df} \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{R}^p \) be a given finite set of \( p \)-component vectors and \( x_k, x_k \in X \). Consider the following expression: \( 1 - \ominus^{\alpha}_{\sigma}\{x_k - x_i\} \). Hence, in accordance with Proposition 3 we can obtain: \( 1 - \min\{1, \sigma \cdot \sum_{i=1}^{p} |x_k - x_i|^\alpha\}^{1/\alpha} \), where \( \sigma \) is the reciprocal of the largest value of the sum \( \sum_{i=1}^{p} |x_k - x_i|\alpha \). i.e. \( \sigma =_{df} 1 / \max\{\sum_{i=1}^{p} |x_k - x_i|\alpha / x_k, x_k \in X; x_i \neq x_k\} \). And finally, the following distance function can be obtained: \( \rho(x_\alpha, x_{\alpha}) =_{df} 1 - \min\{1, \sigma \} \sum_{i=1}^{p} |x_k - x_i|\alpha \}^{1/\alpha} \). It is easily to show that \( \rho(x, x_k) \) equals to the distance function of the Minkowski class.

We observe the one-dimensional fuzzy t-equivalence \( \rho_{\sigma} \), introduced in the previous Example 1, can be interpreted as a particular case of distance function of the Minkowski class. Moreover, the considered in [3] relation \( \rho(x, y) =_{df} \max\{0, 1 - |x - y|\} \) remains a fuzzy t-equivalence for any \( \alpha \geq 1 \). In fact, the following proposition holds.

**Proposition 6**

Let \( \alpha \geq 1 \). Then \( \rho(x, y) =_{df} \max\{0, 1 - |x - y|\} \) is a fuzzy t-equivalence with respect to the generalised Łukasiewicz’s t-norm.

**Proof:**

Assume that \( \alpha \geq 1 \). It is sufficient to show that \( \rho \) is t-transitive, i.e. \( \rho(x, z) \geq \rho(x, y) \ominus \rho(y, z) \) (for any \( x, y, z \in \mathbb{R} \)), where \( \ominus \) is the generalised Łukasiewicz’s t-norm. Equivalently, the following inequality should be shown: \( \max\{0, 1 - |x - z|\} \geq \max\{0, \max\{0, 1 - |x - y|\} + \max\{0, 1 - |y - z|\} - 1\}^{1/\alpha} \). And hence: \( \max\{0, 1 - |x - z|\} \geq \max\{0, \max\{0, 1 - |x - y|\} + \max\{0, 1 - |y - z|\} - 1\} \). Since any absolute value \( |x - z|, |x - y|, \text{ and } |y - z| \) may be greater than, equal to, or less than 1, in general, \( 3^3 = 27 \) cases should be considered (eventually reduced to \( 2^3 = 8 \)). However, the most important is the case when \( |x - z| = 1, |x - y| < 1 \text{ and } |y - z| < 1 \). Hence, the following inequality should be shown: \( (1 - |x - y|)^\alpha + (1 - |y - z|)^\alpha \leq (1 - |x - y|)^\alpha + (1 - |x - y|)^\alpha \). This case is considered below.

Since \( 1 = |x - z| \leq |x - y| + |y - z|, |x - y|, |y - z| < 1 \text{ and } 1 - |x - z| \geq 1 - (|x - y| + |y - z|) \), the above inequality is always satisfied. In fact, for any \( \alpha \geq 1 \) we have: \( (1 - |x - y|)^\alpha + (1 - |y - z|)^\alpha \leq (1 - |x - y|)^\alpha + (1 - |x - y|)^\alpha \).
z \mid + 1)^2 \leq (1 - \mid x - z \mid + 1)^2 = (1 - 1 + 1)^2 = 1^2 = 1 \text{ (since } [\alpha] \leq \alpha \leq [\alpha] \text{, where } [\alpha] \text{ and } [\alpha] \text{ are the corresponding floor and ceiling functions and } a^b + b^b \leq (a + b)^b). \]

4. The generalised fuzzy rough approximations

The concept of the rough set, first introduced in [7], has inspired variety of research of both theoretical and practical nature. The basic idea is that conclusions are drawn with some approximation only and are not exact as in the case of classical logic. It was presented an exact mathematical formulation of the notion of approximative (rough) equality of sets in a given approximation space, understood as a pair $A =_a (U, \rho)$, where $U$ is a certain set called universe and $\rho \subseteq U \times U$ is an equivalence relation. The rough set concept can be of some importance, primarily in some branches of artificial intelligence, such as inductive reasoning, automatic classification, pattern recognition, information systems and decision tables, state identification, learning algorithms, cluster analysis, measurement theory, taxonomy, and so on [8,9].

In general, the lower and upper approximations of a given subset $X \subseteq U$ are computed using $\rho$ and defined as follows: $\overline{A}(X) =_a \{ x \in U / [x] \subseteq X \}$ and $\underline{A}(X) =_a \{ x \in U / [x] \cap X \neq \emptyset \}$, respectively. Equivalently, we have: $\overline{A}(X) =_a \emptyset \cup \{ x \in U \cup [x] \cap X \neq \emptyset \}$ and $\underline{A}(X) =_a \emptyset \cup \{ x \in U \cup [x] \cap X \neq \emptyset \}$. Obviously, the obtained equivalence classes (called also elementary sets or atoms) $[x]_p$ in $U/\rho$ (the quotient set) either coincide or are disjoint.

An equivalent version of the above two approximations was originally proposed in [3]. And so, we have:

$$\begin{align*}
\forall x (\rho y \Rightarrow x \in X) & \iff \exists x (x \in X). \\
\forall x (\rho y \Rightarrow x \in X) & \iff \exists x (x \in X).
\end{align*}$$

By definition, it follows that $y \in \overline{A}(X) \iff [y]_p \subseteq X$ and $y \in \underline{A}(X) \iff [y]_p \cap X \neq \emptyset$. And hence, the following two properties should be satisfied: $[y]_p \subseteq X \iff \forall x (\rho y \Rightarrow x \in X)$ and $[y]_p \cap X \neq \emptyset \iff \exists x (x \in X)$. A more formal treatment is given below.

The following designations are used in the next proofs (the names associated with some primitive and/or derived rules are in accordance with the corresponding Łukasiewicz’s symbols): ‘− $\wedge$’ (rule of omitting a disjunction), ‘− $\wedge$’ (rule of omitting a conjunction), ‘− $\wedge$’ (rule of detachment for implication or omitting an implication), ‘− $\wedge$’ (rule of negating an implication), ‘− $\wedge$’ (rule of negating a conjunction), ‘− $\wedge$’ (rule of substitution for equivalence), ‘− $\wedge$’ ‘− $\wedge$’, ‘− $\wedge$’ and ‘− $\wedge$’ (rules of omitting an universal and an existential bounded quantifiers and also negating an universal and an existential bounded quantifiers, respectively). The introduced abbreviations ‘$\wedge$’, ‘$\wedge$’, and ‘$\wedge$’, denote: assumption(s), assumption(s) of indirect proof, and contradiction, respectively. Provided there is no ambiguity and depending on the context, by ‘$a$’ it is also denoted an element of $U$. Obviously, any element belongs to $U$ and any set of such elements is subset of $U$.

Proposition 7

$[y]_p \subseteq X \iff \forall x (\rho y \Rightarrow x \in X)$

Proof (if-condition):

\begin{enumerate}
\item $[y]_p \subseteq X \quad \text{(a)}$
\item $\forall x (\rho y \Rightarrow x \in X) \quad \text{(aip)}$
\item $\exists x (x \in y) \quad \text{[N\forall\wedge, NC, SR : 2]}$
\end{enumerate}

Proof (only if-condition):

\begin{enumerate}
\item $\forall x (\rho y \Rightarrow x \in X) \quad \text{(a)}$
\item $[y]_p \subseteq X \quad \text{(aip)}$
\item $\forall x (x \in y) \quad \text{[N\forall\wedge, NC, SR : 3]}$
\item $\exists x (x \in y) \quad \text{[N\forall\wedge, NC, SR : 4]}$
\item $\exists x (x \in y) \quad \text{[N\forall\wedge, NC, SR : 5]}$
\end{enumerate}

Proposition 8
\[ y \lor X \neq \emptyset \Leftrightarrow \exists (x \lor y \land x \in X) \]

**Proof (if-condition):**

1. \[ |y|_p \cap X \neq \emptyset \]
2. \[ \sim \exists (x \lor y \land x \in X) \]
3. \[ \forall (x \lor y \land x \in X) \]
4. \[ \exists (x \in [y]|_p \cap X) \]
5. \[ a \in U \]
6. \[ a \in [y]|_p \]
7. \[ a \lor y \]
8. \[ a \in U \Rightarrow a \lor y \land a \neq X \]
9. \[ \forall x \in [y]|_p \cap X \]
10. \[ a \neq X \]
11. \[ \forall x \in [y]|_p \cap X \]

**Proof (only if-condition):**

1. \[ a \in U \]
2. \[ a \in [y]|_p \cap X \]
3. \[ a \neq X \]
4. \[ a \neq [y]|_p \cap X \]
5. \[ a \lor y \]
6. \[ a \lor y \land a \neq X \]
7. \[ a \lor y \land a \neq X \]
8. \[ a \lor y \land a \neq X \]
9. \[ a \lor y \land a \neq X \]
10. \[ a \lor y \land a \neq X \]

Let \( X \subseteq U \). We shall say that \( X \) is exact (or measurable) in \( A \) if and only if \( \Delta(X) = \bar{A}(X) \). And hence, \( X \) is exact in \( A \) if and only if \( X \) is a composed set in \( A \), i.e., a finite union of elementary sets. Any \( X, Y \subseteq U \) are said to be roughly equal (roughly bottom-equal or roughly top-equal) in \( A \) if and only if \( X \) and \( Y \) have the same lower and upper approximations (either the same lower approximations or the same upper approximations) in \( A \). It is easy to show the above notions of ‘roughly equal’, ‘roughly bottom-equal’ and ‘roughly top-equal’ are equivalence relations on \( P(U) \) (the powerset of \( U \)). The corresponding equivalence classes are said to be rough (lower, upper) sets. Therefore, if \( X \) is not exact in \( A \), then \( X \) will belong to some subfamily \( \subseteq P(U) \) called rough set [8].

In accordance to the above considerations, any rough set is related to some ordered pair \((X_1, X_2)\), where \( X_1 = \Delta(X) \) and \( X_2 = \bar{A}(X) \) [3,10]. The proposed here predicate-oriented version for \( \Delta(X) \) and \( \bar{A}(X) \) was extended in the area of fuzzy sets. More exactly, the lower and upper approximations associated with any fuzzy set can be constructed by means of the notions of a fuzzy implication, a t-norm and a fuzzy t-equivalence. And hence, the following lower and upper approximations of a fuzzy set \( \mu \) in \( U \) were introduced [10]. Provided there is no ambiguity and for convenience, here the domain of a fuzzy set \( X \) is denoted by \( U \).

\[ \mu(y) = \text{inf}(x \in U / \rho(x,y) \Rightarrow \mu(x)) \]

\[ \bar{\mu}(y) = \text{sup}(x \in U / \rho(x,y) \Rightarrow \mu(x)) \]

Let now \( \rho \) be a fuzzy t-equivalence in \( U \) and \( y \in U \). The \( \rho \)-forest of \( y \) is the fuzzy set \( py \) having membership function \( \rho y(x) = \text{inf}(x \in U / \rho(x,y) \Rightarrow \mu(x)) \) for all \( x \in U \).

**Example 3**

Consider the fuzzy t-equivalence relation \( \rho_\alpha \) of Example 1. In accordance with Proposition 6, this property of \( \rho_\alpha \) is preserved assuming \( \alpha \geq 1 \). As an example, the \( \rho_\alpha \)-forests of \( x_1 \) and \( x_3 \) are disjoint, i.e. they have an empty \( t \)-intersection. In fact, we have: \( \rho_1 x_1 \cap \rho_1 x_2 \) \( x_1 = \rho_1 x_1 \land \rho_1 x_2 \) \( \rho_1 x_3 \) \( (x_1, x_3) = \max(0, \rho_1 x_3 + \rho_1 x_3 - 1)^{1/\alpha} = 0 \) (for any \( \alpha \geq 1 \) and \( i = 1,2,3,4 \)), e.g. \( \rho_1 x_1 \cap \rho_1 x_2 \) \( x_1 = \max(0, (1/3)^\alpha + (2/3)^\alpha - 1)^{1/\alpha} = 0 \) (since \( 2^\alpha \leq 3^\alpha - 1 \), for \( \alpha \geq 1 \)). The corresponding rows for \( x_1 \) and \( x_3 \) in \( M_\rho \) have all elements different.

Let now consider the \( \rho_\alpha \)-forests of \( x_1 \) and \( x_3 \). Since \( \rho_1 x_3 (x_2) = \rho_1 x_3 (x_2) = 2/3 \) then \( \rho_1 x_1 \cap \rho_1 x_2 \) \( x_2 = 2 \) belongs to degree 1/3 to the \( t \)-intersection of the above two \( \rho_\alpha \)-forests, i.e. they are not disjoint.

Let \( \rho \) be a fuzzy relation that models an approximate equality. Then, we shall say that \( py \) is a fuzzy similarity class of \( y \). According to the last example, an element \( y \in py \) can also belong to other, different similarity classes to a certain degree. In fact, the following list of candidate definitions for the lower (the upper) approximation of \( \mu \) should be considered [3].

Any \( y \in U \) belongs to the lower (the upper) approximation of \( \mu \) to the degree to which:
All fuzzy similarity classes containing \( y \) are included in \( \mu \) (have a nonempty intersection with \( \mu \)).

b. At least one fuzzy similarity class containing \( y \) is included in \( \mu \) (has a nonempty intersection with \( \mu \)) and

c. The fuzzy similarity class \( \mu(y) \) is included in \( \mu \) (has a nonempty intersection with \( \mu \)).

In accordance with the above considerations, the notions of tight, loose and usual lower and upper approximations were introduced. For convenience, the following designations for tight, loose and usual lower (upper) approximations are used below: \( \underline{\mu} \), \( \underline{\mu} \), and \( \mu \) (\( \mu \), \( \mu \), and \( \mu \)), respectively. In the case of usual approximations the index 'u' may be omitted. The following extended versions can be proposed (for all \( y \in U \)).

**The tight, loose and usual lower approximations:**

\[
\underline{\mu}(y) =_{df} \inf\{z \in U / \rho(z(y)) \Rightarrow_{\alpha} \inf\{x \in U / \rho(z(x)) \Rightarrow_{\alpha} \mu(x)\}\},
\]

\[
\mu(y) =_{df} \sup\{z \in U / \rho(z(y)) \Rightarrow_{\alpha} \inf\{x \in U / \rho(z(x)) \Rightarrow_{\alpha} \mu(x)\}\},
\]

\[
\mu(y) =_{df} \inf\{x \in U / \rho(y(x)) \Rightarrow_{\alpha} \mu(x)\}.
\]

The tight, loose and usual upper approximations:

\[
\tilde{\mu}(y) =_{df} \inf\{z \in U / \rho(z(x)) \Rightarrow_{\alpha} \sup\{x \in U / \rho(z(x)) \Rightarrow_{\alpha} \mu(x)\}\},
\]

\[
\tilde{\mu}(y) =_{df} \sup\{z \in U / \rho(z(y)) \Rightarrow_{\alpha} \sup\{x \in U / \rho(z(x)) \Rightarrow_{\alpha} \mu(x)\}\},
\]

\[
\tilde{\mu}(y) =_{df} \sup\{x \in U / \rho(y(x)) \Rightarrow_{\alpha} \mu(x)\}.
\]

According to the last definitions, \( \Rightarrow_{\alpha} \) and \( \Rightarrow_{\alpha} \) denote the generalised Lukasiewicz’s t-norm and fuzzy implication (Proposition 4, Corollaries 1 and 2), respectively.

**Example 4**

Let \( U \) be the subset \( Y \) from Example 1 and \( \rho(x) \) be the obtained fuzzy t-equivalence represented by \( \theta(x) \). Assume that \( \mu =_{df} (3/5, 4/5, 1/5, 2/5) \) is a fuzzy set defined in \( Y \). In accordance with the above definitions, e.g. the following usual lower and usual upper approximations are obtained (the index 'u' is omitted here).

\[
\begin{array}{|c|c|c|}
\hline
\alpha & U & \tilde{\mu} \\
\hline
1 & (3/5, 8/15, 1/5, 2/5) & (3/5, 4/5, 1/5, 2/5) \\
2 & (3/5, \sqrt{34}/15, 1/5, 2/5) & (3/5, 4/5, \sqrt{19}/15, 2/5) \\
\hline
\end{array}
\]

As an illustration, the computations related to \( \hat{\mu}(x_3) \) and \( \mu(x_3) \), i.e. \( y =_{df} x_3 \) and \( \alpha =_{df} 2 \), are given below. And so, in the case of the usual lower approximation we can obtain:

\[
\rho(x_1, x_3) \Rightarrow_{\alpha} \mu(x_1) = 1/3 \Rightarrow_{\alpha} 3/5 = \min\{1, 1 - (1/3)^2 + (3/5)^2\} = 1, \\
\rho(x_2, x_3) \Rightarrow_{\alpha} \mu(x_2) = 2/3 \Rightarrow_{\alpha} 4/5 = 1, \\
\rho(x_3, x_1) \Rightarrow_{\alpha} \mu(x_3) = 1 \Rightarrow_{\alpha} 1/5 = 1/5, \\
\rho(x_3, x_3) \Rightarrow_{\alpha} \mu(x_3) = 2/3 \Rightarrow_{\alpha} 2/5 = \sqrt{161}/15.
\]

And hence: \( \hat{\mu}(x_3) = \min\{1, 1, 1/5, \sqrt{161}/15\} = 1/5 \) (since \( 3^2 < 161 \)).

In a similar way, in the case of the usual upper approximation we have:

\[
\rho(x_1, x_3) \Rightarrow_{\alpha} \mu(x_1) = 1/3 \Rightarrow_{\alpha} 3/5 = \max\{0, (1/3)^2 + (3/5)^2 - 1\} = 1/2 = 0, \\
\rho(x_2, x_3) \Rightarrow_{\alpha} \mu(x_2) = 2/3 \Rightarrow_{\alpha} 4/5 = \sqrt{19}/15, \\
\rho(x_3, x_3) \Rightarrow_{\alpha} \mu(x_3) = 1 \Rightarrow_{\alpha} 1/5 = 1/5, \\
\rho(x_3, x_3) \Rightarrow_{\alpha} \mu(x_3) = 2/3 \Rightarrow_{\alpha} 2/5 = 0.
\]

Therefore: \( \hat{\mu}(x_3) = \max\{0, \sqrt{19}/15, 1/5, 0\} = \sqrt{19}/15 \) (since \( 3^2 < 19 \)). We observe a better approximation using \( \alpha = 2 \), i.e. the obtained Hamming distance: \( d(\mu, \mu) =_{df} \sum_{x \in X} |\mu(x) - \hat{\mu}(x)| \) is less than this one for \( \alpha = 1 \) (the classical case). And so: \( \mu \mid_{\alpha = 1} \subseteq \mu \mid_{\alpha = 2} \subseteq \mu \subseteq \tilde{\mu} \mid_{\alpha = 2} \subseteq \tilde{\mu} \mid_{\alpha = 1} \).

**Proposition 9**

\[
\mu \mid_{\alpha = 1} \subseteq \mu \mid_{\alpha = 2} \subseteq \mu \subseteq \tilde{\mu} \mid_{\alpha = 2} \subseteq \tilde{\mu} \mid_{\alpha = 1}
\]

**Proof:**

It is sufficient to show that: (a) \( \min\{1, 1 - x + y\}^2 \leq \min\{1, 1 - x^2 + y^2\} \) and (b) \( \max\{0, x + y - 1\}^2 \geq \max\{0, x^2 + y^2 - 1\} \).

a) Let \( x \leq y \). Hence: \( x^2 \leq y^2, 1 - x + y \geq 1, 1 - x^2 + y^2 \geq 1 \) and the left (L) and right (R) sides coincide, \( L = R = 1 \).

Assume now that \( x > y \). We have: \( L = (1 - x + y)^2 \) and \( R = 1 - x^2 + y^2 \). It is necessary to
show that: \((1 - x + y)^2 \leq 1 - x^2 + y^2\), \((x, y \in [0, 1], x > y)\). And so, we have: \((1 - x + y)^2 \leq 1 - x^2 + y^2\), iff \((1 + y - x)^2 \leq 1 + y^2 - x^2\), iff \(1 + 2y - 2x + y^2 - 2y^2 + x^2 \leq 1 + y^2 - x^2\), iff \(2y - 2x - 2y^2 + x^2 \leq 0\), iff \(y - x - (y - x) \leq 0\), iff \((y - x)(1 - x) \leq 0\). Since \(x > y\) and \(x \leq 1\) then \(y - x < 0\) and \(1 - x \geq 0\). Hence, the last inequality is always satisfied.

b) Let \(x + y \leq 1\). Since \(x + y - 1 \leq 0\) the left side \(L = 0\). Also \((x + y)^2 \leq 1\) and hence \(x^2 + y^2 \leq 1\) (for \(x, y \geq 0\), for \(x, y \leq 0\)). Then: \(R = 0\).

Let now \(x + y > 1\). Since \(x + y - 1 > 0\) then \(L = (x + y - 1)^2 > 0\). And hence, it is sufficient to show that \((x + y - 1)^2 \geq x^2 + y^2 - 1\). We have: \((x + y - 1)^2 \geq x^2 + y^2 - 1\) iff \((x + y - 1)^2 \geq x^2 + y^2 - 1\) iff \(x^2 + y^2 + 2xy - 2x - 2y + 1 \geq x^2 + y^2 - 1\) iff \(2x - 2y - 2y^2 + 2x^2 \geq 0\). Since \(x > y\) and \(x > 1\) then \(x = 0\) will implicate \(y > 1\) (contra \(y \leq 1\)). Hence \(x \neq 0\). Similarly \(y \neq 0\) and \(xy > 0\). On the other hand \(x, y \leq 1\). And so, \(0 < x + y < 2\). Hence, \(1 + xy \geq x + y\) is always satisfied.

Obviously, the above inclusions are satisfied for any \(\alpha \geq 1\). The corresponding proofs assuming \(x \leq y\) (assuming \(x + y \leq 1\)) are trivial: we have \(L = R\), e.g. case (b): since \(x + y \leq 1\) then \(L = 0\). From \(x \geq x^a\) and \(y \geq y^a\) it follows that \(1 \geq x + y \geq x^a + y^a\). Hence \(x^a + y^a \geq 1 - 0\) and \(R = 0\).

The proofs related to \(x > y\) or \(x + y > 1\) (cases (a) and (b), respectively) correspond to the following two inequalities: \((1 - x + y)^a \leq 1 - x^a + y^a\), \((x + y - 1)^a \geq x^a + y^a - 1\) (for \(x > y\)) and \((x + y - 1)^a \geq x^a + y^a - 1\) (for \(x + y > 1\)).

Let consider case (a). Since \(y < x\) then \(y - x > 0\). Hence \(1 + y - x < 1\) and \((1 + y - x)^a \leq 1 + y - x < 1\). Similarly, \(y^a < x^a, y^a < 0\) and \(1 + y^a - x^a < 1\). Since \(x^a \geq y^a\) then for any increasing \(\alpha \geq 1\) the absolute value \(|y^a - x^a|\) will be decreasing and hence \(1 - |y^a - x^a| = 1 - x^a + y^a\) will be increasing. At the same time the left side is decreasing (see the example table below): util \(x = 1 / 2, y = 1 / 3\), here e.g. \(R(4)\) is about twice greater than \(L(4)\).

Consider now case (b). Since \(x + y > 1\) then \(x, y \neq 0\) and the obtained value for \(x^a + y^a\) may or not be less than 1 (depending on \(\alpha\)). Let \(x^a + y^a \leq 1\). Then \(R = 0\) and \(L = (x + y - 1)^a \geq 0\) since \(x + y - 1 \geq 0\). Otherwise, should be satisfied the following equivalent inequality: \((x + y - 1)^a + 1 \geq x^a + y^a\). Since \(0 < x, y \leq 1\) then \(0 < x + y \leq 2\). According to the last inequality, assuming \(x = y = 1\) we have: \(2 = 2\). In any other situation the left side will not be less than the right one. In fact, for any increasing \(\alpha \geq 1\) we can obtain: \(L(\alpha) \leq R(\alpha)\), case (a) and \(L(\alpha) \geq R(\alpha)\), case (b).

A more formal treatment is omitted.

<table>
<thead>
<tr>
<th>(\alpha)</th>
<th>(L(\alpha))</th>
<th>(R(\alpha))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1080 / 1296</td>
<td>1080 / 1296</td>
</tr>
<tr>
<td>2</td>
<td>900 / 1296</td>
<td>1116 / 1296</td>
</tr>
<tr>
<td>3</td>
<td>750 / 1296</td>
<td>1182 / 1296</td>
</tr>
<tr>
<td>4</td>
<td>625 / 1296</td>
<td>1231 / 1296</td>
</tr>
</tbody>
</table>

5. Conclusions

Any new t-norm implies some new applications, e.g. such as: introduction of new t-norm based measures or also computations related to the possibility of fuzzy events, specification of new commutative and associative copulas, new possibilities to combine criteria in multicriteria decision making (for evaluation the truth degrees of compound formulae), new kind of fuzzy t-equivalence and so on. Fuzzy rough sets have become an important part of modern computer science. It has presented a possibility of generalisation of the notions of lower and upper approximations used in fuzzy rough sets and also of obtaining better such approximations. More formally, the obtained Hamming distance \(d(\mu, \tilde{\mu})\) is decreasing with respect to increasing \(\alpha \geq 1\). The so-obtained approximations can be used in very many areas, e.g. such as medical imaging, fuzzy control, data bases, and so on. Any such applications may be topics for further research.

References


Iwan Tabakow
Dr. Iwan Tabakow received his M.Sc., Ph.D. and D.Sc. degrees as follows: Wroclaw University of Technology (M.Sc., Computer Engineering 1969), The University of Wroclaw (M.Sc., Mathematics 1972), Wroclaw University of Technology (Ph.D. and D.Sc., Computer Science, 1972 and 1975, respectively). His professional experience includes: assistant professor and lecturer (Institute of technical Cybernetics, Wroclaw University of Technology and the Department of Mathematics, The University of Wroclaw: 1973 - 1975), assistant professor, associate professor, and full professor at the Department of Computer Science and Engineering, Technical University of Sofia (1976 - 1984, 1984 - 1996, and 1996 - 1998, respectively), full professor in Computer Science at the Institute of Informatics, Faculty of Computer Science and Management (Wroclaw University of Technology: 1998 - 2014). Dr. Tabakow is currently a lecturer in the following three topics: "Discrete mathematical structures", "Fault Diagnosis of Digital Systems", and "Petri Nets". Areas of main research interests: discrete structures, formal logic, system diagnosis and Petri nets. He has about eighty publications all in the above given areas (including four monographs, journal and refereed conference papers). He was founder, chairman, and/or reviewer of a number of national and international scientific conferences, external examiner, and so on. He is member of Gesellschaft für Informatic e.V. Bonn.
Face Detection Approach Based on Fuzzy Logic and GST

Arash Kiani Zain Abad¹, Mostafa Asghari Pile Savar² and Vahid Rostami³

¹ Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University,Qazvin,Iran  
A.Kiani@Qiau.ac.ir

² Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University,Qazvin,Iran  
MostafAsghary@gmail.com

³ Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University,Qazvin,Iran  
Vh_Rostami@yahoo.com

Abstract
This paper proposes a fast and confident approach to face detection. Initially we detect and segment the regions of image with human skin color. The regions with similar features are merged in order to generate a bigger region. Following the extraction of each region's features, the resulted data is given to the fuzzy system trying to determine regions which are most likely to belong to skin. After performing preprocessing, the regions with highest possibility of containing face are examined and similar regions are merged to generate a large region. After segmenting the face regions of image, the regions are divided into two categories of face and non-face by applying the Generalized Symmetry Transform (GST) algorithm. The results prove that new algorithm not only runs relatively fast but also provides an improved performance.

Keywords: Face detection, edge detection, fuzzy logic, Generalized Symmetry Transform (GST), Image segmentation.

1. Introduction

The human face processing algorithms such as face feature extraction and face detection are of crucial importance and have various applications. Lots of applications depend on the size and position of the face in the image. Most of the real-world applications rely on face size and position estimation methods. Generally two issues should be considered in face detection area: detection rate and speed; which are closely interrelated. Different algorithms are attempting to keep a balance between speed and calculation time. Yang et al. [1] and [2] are admirable instances of old methods with optimized calculation time. As another example, Gatokomal is a face detection system which is capable of real-time face detection in colorful video streams. In this system, after segmenting face color, PCA [3] is incorporated into determination of whether a specific region belongs to face skin or not.

The [4] was a background for AdaBoost which used the inseparable face concept for face detection. They proposed two-class AdaBoost training algorithm for efficient classification training and cascading for elimination of non-face images.

The [5] proposed a face detection approach in grayscale images which detects regions similar to eye which are the same size as real eyes with darker color compared to surrounding regions. Assuming there are eyes in the given face, a pair of eye regions (eye scalable) is selected. In case of matching the position of selected regions with human face’s characteristic, they are declared as eyes.

The [6] after adjusting the color of input images and crooked face; straightened the skewed corners of the mouth and determined the discriminating function for specifying the position of the eyes. The [7] and colleagues applying the conditional distribution of brightness of the skin color information and using the image correction process, extracted and drew a modeled rectangle around the skin area. Finally they applied an adaptation framework by means of linear transformation in each rectangular area of skin for face detection.

Classifier neural network is widely used in face detection, mostly while face is in the center of the image.

Rowley [8] and his colleagues presented a neural network-based face detection system, a corneal connection (depending on the retina) neural network to verify the small windows of the image and decide which window contains human face. Mansour [9] et al proposed a face detection algorithm based on light control, techniques for detection and classification of skin color. Their approach detects the face rectangle that contains eyes and mouth and areas coming from skin detection process using color segmentation. Then it analyzes all features that might be the face (eyes and mouth) and ensures the determination of rectangle containing eyes and mouth by using a neural network.

Commonly used cascade-based face detection techniques such as the ones given by Viola and Jones [13] only coarse face detection results.
Aforementioned algorithm works without usage of skin color and different pixel in successive frames. In fact it grids the image into blocks including 32*32 pixels, (Considering the fact that the smaller blocks results in higher computational cost). Each block is evaluated using NN based method. This algorithm is fast enough; nevertheless it generates high number of false positive results.

In [15] face detection process is divided into two steps, which are named root and parts step respectively. It is also a cascade based algorithm. In first step a linear SVM is applied in order to detect eyes and mouth using frequency features. The sizes of blocks are 72*60 pixels. In second step an SVDD (support vector data descriptor) is used in order to distinguish face from non-face regions with blocks in size of 36*30 blocks. Even though the second step is slower than the first one, yet the reduction of the number of blocks in the first step, improves the overall performance of the algorithm. The performance of the method is defendable only in high-resolution images with faces bigger than 80*80 pixels.

The algorithm described in [14] incorporates the color of face skin. The aforementioned step is followed by thresholding method in outdoor and indoor environments using the HSV color space. The resultant image is converted into binary image which makes it simple to distinguish regions with face skin color from other ones. The binary image is enhanced to benefit next step which applies Point of Interest to count the interesting points. If the number of interesting points inside a region falls between the minimum and maximum threshold, it is detected as face with indication of a Bounding Box.

In [10] the authors illustrate their motivation of using the Generalized Symmetry Transform (GST) for detected location of eyes and mouths in image. They are extracting the features from image and using SVM to find skin section of picture. They also applied motion detection and then used horizontal adjustment to find symmetry of segmented region. At the end, they used a GST algorithm to verifying that whether the region is a face or not. A lot of research work has been done in the field of face detection. Some areas of study have attempted to convert images into smaller windows separated by application of GST algorithm to decide whether or not window contains a face. However, the above calculations in the number of regions and issues of adaptive network topology have limited its use. In some cases, these problems may result in high complexity. For example, if we use the neural network to find the faces in the input images without any prior knowledge, the calculation will be very heavy and will take long.

And also when using a GST [10] without any pre-processing, it takes longer to find symmetry of each region in the image.

The basic problem is to reduce the calculation cost by pre-processing steps. The number of regions given to the symmetry system, to a great extent will lead to less computation. This pre-classification is of considerable importance because if the rule base is too complex, some of the faces in the image might be ignored.

In our approach, we delete some non-face regions using an inference engine developed by a set of flexible rules and also applying a few simple and reliable features. Consequently, number of regions given to symmetry system is decreased.

The aforementioned fuzzy system has removed some of non-face regions, so number of regions proposed to symmetry system has decreased. Figure 1 shows the block diagram of the proposed system. As it is illustrated, the input of the system is provided with color images, and after segmentation the color space is divided into skin and non-skin segments.

In the pre-classification step, a small number of simple and reliable features are being extracted for fuzzy inference engine.

If the result of fuzzy inference engine indicates the existence of face in the current region, so the region is sent to symmetry system for making final decision.

The output of system displays the detected face in the image. In section II, the image segmentation is described. In section III, the proposed fuzzy system is described. The facial symmetry system is explained in Section IV. Our experimental results are presented in Section V and finally Section VI provides the conclusion.

![Figure 1: The presented face detection approach](image)

### 2. Face Region Detection

Firstly, some regions of color images which are the input of the system should be removed to reduce the search space. By using Bayesian methods, the skin or skin like regions are separated from non-skin ones. To model skin color, skin samples from 50 thousand randomly
selected individuals from various ethnic groups are provided through internet. Then, using Chromatic color space, the light intensity can be removed through normalization process with the following formula:

\[ r = \frac{R}{R + B + G}, g = \frac{G}{R + B + G}, b = \frac{B}{R + B + G} \]  

(1)

Figure 2 (left) shows the distributed skin samples of the Chromatic color space. This distribution could be modeled based on Gaussian distribution with following parameters:

\[ m = E(x), c = E((x, m)(x, m)^T), x = (r, b)^T \] 

(2)

Where m is mean, and c is the covariance and vector x is a value of r and b.

Using Mahalanobis measurement, regions of the image having a color closer to the color of the skin are identified which is shown in formula 3.

\[ S(r, b) = \exp[-0.5 (x - m)^T c^{-1}(x - m)] \] 

(3)

The output of S is between 0 and 1 that shapes a binary image. You can see the result in figure 3.

After this step, the distance of each region from other ones is examined and if this distance is less than a threshold value (10 pixels); these regions are merged into each other.

3. Design of Fuzzy Inference System

Image Segmentation in the previous step, separated the skin and non-skin regions from each other. Now we need a fuzzy system to classify these regions as face or non-face. In order to reduce the computation of the system, a fuzzy inference engine, which we call it the pre-classifier, is used. First, we extract some features from each region. At the next level, a fuzzy inference engine based on output of the previous step determines that which regions are more likely to be face.

3.1. Region features extraction

First, the regions received from classification step are converted into binary regions. If the area of the region is smaller than fifteen pixels, it will be removed. Features involved in the decision are: the number of holes in one region and the ratio of length to width of the region.

3.2. Pre-classification: a fuzzy inference engine

After extraction of relevant features, the data will be given to fuzzy inference engine to make final decision on whether or not this region contains face. You can see the rules in table 1.

4. Human Face Verification

4.1 Localization of human face based on the geometric features

The aim of facial feature localization is to locate eyes and mouth. Because the features of nose are not robust, we do not localize nose. With the locations of eyes and mouth, human face can be more precisely localized. Generalized Symmetry Transform (GST) [11] is a method to describe symmetry of points. Since the centers of eyes always have
the highest symmetry in human face, we can use GST to locate eyes. Because this method only uses the biometrics distribution features of human face, it is more robust than some other methods under the variation of illumination, pose and expression.

In our proposed method, Firstly we applied color analyze and fuzzy interference engine based on number of holes and height and width proportion in order to eliminate regions which are least likely to be face regions. Aforementioned pre-classification processes are followed by GST based algorithm proposed in this paper.

Table 1: Fuzzy inference engine rules

<table>
<thead>
<tr>
<th>If</th>
<th>Then</th>
<th>Is Face?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hole</td>
<td>Height/Width</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>L</td>
<td>F</td>
</tr>
<tr>
<td>L</td>
<td>M</td>
<td>T</td>
</tr>
<tr>
<td>L</td>
<td>H</td>
<td>F</td>
</tr>
<tr>
<td>M</td>
<td>L</td>
<td>F</td>
</tr>
<tr>
<td>M</td>
<td>M</td>
<td>T</td>
</tr>
<tr>
<td>M</td>
<td>H</td>
<td>F</td>
</tr>
<tr>
<td>H</td>
<td>L</td>
<td>F</td>
</tr>
<tr>
<td>H</td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>H</td>
<td>H</td>
<td>F</td>
</tr>
</tbody>
</table>

4.2. Algorithms of locating eyes and lips

GST and its extended methods, such as Direction Symmetry Transform and Discrete Symmetry Transform, have already been used in analyzing face images [12]. But the existing methods are computationally time-consuming. The reason of this trouble is that different dimensional factors will detect different symmetry centers. In another word, a specific dimension factor can only detect symmetry centers whose dimension is similar to it. So if we use the size of eyes as σ, we can find the centers of eyes. Because no dimension information is known in these existed methods, they have to compute in a large range of dimension. But in our method, the size of eyes can be estimated according to the face detection results, so it only needs to compute in a small range of dimension. Experiment results show that the suitable range is 1/12–1/10 of the width of face image.

The following three steps can localize eyes in a grayscale face image. Firstly, we use Canny operator to detect edges of the gray image (see Fig. 6(b)). The processing result is an edge image (see Fig. 4(b)). Secondly, the symmetry of every point in the edge image is calculated by GST (see Fig. 7). You can see that the best thresholding value is 0.5 for this dataset. In the process of calculating symmetry of every point, there is a problem like the bound effect in convolution. It is how to calculate the symmetry of points whose distances to the bound of the face image are smaller than σ. Since eyes seldom locate in this boundary region, we can set 0 as the symmetry of these points and do not need to calculate their symmetry. Thirdly, 4–7 points with the highest symmetry are selected to locate the centers of eyes. The exact number of selected points is determined based on experimental experience. If too few points are selected, they may locate in the region of one eye. If too many points are selected, some points with high symmetry, which are not in the regions of eyes, may also be selected.

5. Experimental Result

To compare the proposed fuzzy techniques and methods that are only using the GST algorithm; hundred, two hundred, three hundred and four hundred images were randomly selected from the dataset of the University of Massachusetts\(^1\). You can see the sample of image in Figure 9. As expected the obtained results indicated that

\(^1\) http://vis-www.cs.umass.edu/lfw/#download
the FGST algorithm is more accurate and faster than aforementioned solutions. As observed in Figures 8 and 9, fuzzy symmetry detection algorithm dramatically saves the computing time and provides better performance than a conventional symmetric algorithm.

As illustrated in Figure 9, even though the Viola-Jones et al. [13] is relatively fast, yet it generates considerable number of false positive results which is considered as the main drawback of the algorithm. The Nur Baiti Zahir et al [14] enjoys an acceptable performance, however it does not have such a high accuracy, since it merely applies thresholding and POI methods in order to detect faces. Finally the Hakan Cevikalp et al [15] is compared with other algorithms. It uses cascade based methods which results in searching all blocks of the image in different angels. The search process makes the whole system slow. Also it might not succeed in detecting all face blocks of the image.

6. Conclusions

In this paper, according to the high importance of face detection, a fast and efficient system for face detection is proposed. For this matter, first, the regions of the image that have the same or similar color to the human face skin color are selected. Then, using a fuzzy algorithm, we attempted to estimate regions of which the number of pixels and the ratio of length to width are similar to human face characteristic. Finally, with GST algorithm the non-face regions are removed.

The results indicate that are proposed method has good performance and acceptable output. In future works, the implementation of other methods such as context-based approaches are proposed. Also integration of social network and face detection might be interesting topic for discussion.

References


Novel Approach for Optimal Sizing of Stand-alone Hybrid Photovoltaic/Wind Systems Using Evolutionary Algorithms

Zakieh Tolooi1*, Hadi Zayandehroodi2, Alimorad Khajezadeh3
1, 2, 3 Science and research, Islamic Azad University, Kerman, Iran
1Zakiye_tolooe@yahoo.com, 2H.zayandehroodi@yahoo.com, 3Ali.khajezadeh@yahoo.com

Abstract
Nowadays using of new energies in the form of dispersed resources in the worlds is wide spreading. In this article we will design a dispersed production source in the form of a solar/wind hybrid power plant in order to supply the energy of a residential unit according to a sample load pattern based on evolutionary algorithms such as GSA and PSO algorithms to optimize. The aim of aforementioned design is to reduce its costs in a period of 20 years. In order to optimize system costs we will use a new algorithm which is based on collective intelligence namely gravitational search algorithm and also we will use particle swarm optimization algorithm. Finally we can conclude that with an appropriate design of dispersed production resources we will be able to effectively reduce costs and make renewable energy usage more economically.

Keywords: Energy Hybrid Systems, GSA, Optimal Sizing, PSO, Renewable Energies.

1. Introduction
Today new energy usage in different regions of the world is rising. But with respect to the low efficiency of these kinds of energy production units, in many cases we cannot economically justify their usage. One way for solving this problem is to increase the efficiency with the help of multiple sources in the form of hybrid power plants for energy supplying in subsystems and independent loads. Because of PV specifications and WG specifications, solar - wind hybrid unit is one of the usual hybrid power plants. In Fig. 1 we can see a diagram of our hybrid power plant [1]. Outputs of energy production units are attached to a dc bus bar which with the help of a battery, is used for battery bank charging. Parallel to the battery bank we have an inverter (dc/ac convertor) which supplies load required energy. Energy surplus saves in the battery bank with the help of energy production units and when the energy is not enough with respect to load demand, saved energy will be used. One of the most important things in designing of such systems is reliable load supply in subsystems at different weather conditions which is considered as the main constraint [2]. In our problem, the number and types of needed equipment for load supplying with minimized cost is the main aim. System costs include purchasing costs, installation costs and maintenance costs in a period of 20 years. With respect to the nature of the presented model, problem parameters include number and types of solar cells, number and types of wind power plants, number and capacities of battery banks and the type of charger battery. Other parameters of the model such as anemometry data, radiation data, load demand and other information are considered as inputs for the problem. In recent years many researches are established about solar - wind hybrid power plant designing and optimization. In [1] with the help of GA or generic algorithm sizing problem of a solar - wind hybrid system is optimized. Also in [3] sizing of such system is performed with reliability criterion and with the help of PSO algorithm. In some studies energy transforming network is considered in order to minimize the loss [4]. In [5] we have an algorithm for optimized modelling and arrangement of a solar- wind hybrid power plant. Also in [6] and [7] we have a good review of solar/wind hybrid system designing and modelling and after that, some recommendations and suggestions are presented in order to guideline future studies. In recent years using of intelligent algorithms for optimization problems is increasing significantly. Our study in this article because of its nature also shows the ability of these algorithms and altogether, relative studies represent their success in optimization activities.

Fig. 1. Block diagram of the studied hybrid system
For example hybrid power plant designing problem is optimized by GA [1], PSO [2] [3] [8] and Markov chain [9]. So in this article, we will use a new method which is based on collective intelligence in order to optimize wind/solar hybrid power plant problem. In suggested method, we will use gravitational search algorithm or GSA which employs gravity rule and object interactions. For using this algorithm to optimize the problem, variable parameters (number of equipment) are considered as problem dimensions. For model simplification we will waive the type of equipment. So we have 3 variables which are: number of solar cells, number of wind power plants and number of batteries needed for battery bank. In each iteration, problem variables are determined with respect to system constraints.

We will evaluate the effectiveness of our proposed method with using anemometry data and solar radiation data in one area of Ardebil and Mashhad province in which, our systems are used for satisfying load demand of a rural region with 1 kW load pick and 10kW load for 20 years period of time. Moreover this problem is also solved by PSO and the results are compared. In this article we first use sys fluent software for Kite simulation and calculate resultant force, then we use MATLAB software and convert the resultant force to moment and then calculate useful power. In next sections we will explain the procedure and make a comparison in different wind speeds.

2. Modelling and Simulation of System Performance

In this section we will survey system main relations. Also for system simulation, each year is represented by an hourly period of time and simulations all are ran in these periods.

2.1. Solar Cells

With respect to Fig. 2, maximum output power of a solar cell in the day and hour can be calculated by (1):

\[
P^i_P(t) = N_s N_P V^i_{OC}(t) I^i_{SC}(t) FF^i(t) \tag{1}
\]

\[
I^i_{SC}(t) = [I^i_{SC,STC} + K_i (T^i(t) - 25^\circ C)] \frac{G^i(t)}{1000} \tag{2}
\]

\[
V^i_{OC}(t) = V^i_{OC,STC} - K_v T^i(t) \tag{3}
\]

\[
T^i(t) = T^i_{STC} + \frac{N_{COT} - 20^\circ C}{800} G^i(t) \tag{4}
\]

In which \( N_s \) and \( N_p \) are numbers of cells in serial and parallel forms respectively. Also \( V^i_{OC}(t) \) is the open circuit voltage (v), \( I^i_{SC}(t) \) is the open circuit current (A), \( FF^i(t) \) is fill factor, \( I^i_{SC,STC} \) is the open circuit current in standard state (A), \( K_i \) is the thermal coefficient of open circuit current (A/°C), \( G^i(t) \) is the amount of radiation absorbed by cell area (W/m^2), \( V^i_{OC,STC} \) is the open circuit voltage in standard state (v), \( K_v \) is the thermal coefficient of open circuit voltage (V/°C), \( T^i_{STC} \) is the temperature (°C) and NCOT is the nominal temperature of cell performance(°C). Serial cell numbers can be calculated as:

\[
N_s = \frac{V^m_{DC}}{V^m_{OC}} \tag{5}
\]

In which \( V^m_{DC} \) is the maximum input voltage to charger batteries and \( V^m_{OC} \) is the maximum open circuit voltage of solar cells.

![PV module current–voltage and power–voltage characteristics](image)

Fig. 2. PV module current–voltage and power–voltage characteristics

2.2. Battery Charger Batteries

Power transferred from a solar cell to a battery bank can be calculated as:

\[
n_s = \frac{P^i_{PV}(t)}{P^i_{DC}(t)} = n_1 n_2 \tag{6}
\]

In which \( n_s \) is charger battery conversion factor, \( P^i_{PV}(t) \) is the real power transferred from solar cell, \( P^i_{DC}(t) \) is the maximum power available from solar cell, \( n_1 \) is the efficiency of power electronic equipment and \( n_2 \) is the conversion faction which is dependent on battery charge algorithm [2]. The number of charger batteries is equal to total number of solar cell block and can be calculated as (7), In which \( N_{PV} \) is the total number of solar cells, \( P^m_{PV} \) is the
maximum output power of a solar cell and \( P_{ch}^m \) is nominal power of charger battery.

\[
N_{PV}^{ch} = \frac{N_{PV} \cdot P_{PV}^m}{P_{ch}^m}
\]  

(7)

### 2.3. Wind Unit

As shown in Fig. 3, output power of a wind unit is related to wind speed and can be formulated as:

\[
P_{WG}^i(t) = P_1 + \left[ v^i(t) - v_1 \right] \frac{P_2 - P_1}{v_2 - v_1}, \quad v_1 < v^i(t) < v_2
\]

(8)

In which \( P_{WG}^i(t) \) is the delivered power to battery bank from wind turbine in \( i \)th day and \( i \)th hour (W) and \((P_1,v_1)\) and \((P_2,v_2)\) are (wind speed and wind turbine power) couples with respect to tables available in [2]. The amount of wind speed is proportional to turbine installation height and can be calculated as:

\[
v^i(t,h) = v_{ref}(t) \left( \frac{h}{h_{ref}} \right)\alpha
\]

(9)

In which \( v^i(t,h) \) is the wind speed at turbine installation height, \( v_{ref}(t) \) is the amount of reference wind and \( h_{ref} \) is the amount of reference height. Also \( \alpha \) is a constant (named power low constant) and its amount varies between 1/7 and 1/4.

![WG power versus wind speed characteristic](image)

**Fig 3.** WG power versus wind speed characteristic

### 2.4. Battery Bank

With respect to battery banks, the bank which is used in this model can only be discharged only 80 percent. This amount is related to discharge depth which is determined by system designer. Minimum allowable battery capacity in the discharging process can be formulated as:

\[
C_{min} = (1 - DOD) \cdot C_n
\]

(10)

In which \( C_{min} \) is the minimum allowable battery capacity, \( DOD \) is the maximum discharge depth and \( C_n \) is nominal battery capacity. The amount of battery capacity is proportional to time and changes in the research period as:

\[
C_i(t) = C_i(t-1) + n_B \frac{P_i(t)}{V_{BUS}} \Delta t
\]

(11)

\[
C_i(24) = C_i^{1+1}(0)
\]

(12)

In which \( C_i(t) \) and \( C_i(t-1) \) are available battery capacity (ah) at \( t \)th and \( t-1 \)th day respectively, \( n_B \) is battery efficiency , \( P_i(t) \) is battery input/output power , \( V_{BUS} \) is bus bar voltage (dc) and \( \Delta t \) is the simulation time step. The number of serial batteries can be calculated as:

\[
n_B^s = \frac{V_{BUS}}{V_B}
\]

(13)

In which \( n_B^s \) is the number of serial batteries and \( V_B \) is the nominal voltage of each battery.

### 2.4. Complete Model

The total amount of transferred power from solar cells and wind units can be calculated as:

\[
P_i^i(t) = N_{PV} \cdot P_{PV}^i(t) + N_{WG} \cdot P_{WG}^i(t), \quad 1 \leq i \leq 365, \quad 1 \leq t \leq 24
\]

(14)

In which \( N_{PV} \) is the total number of solar cells and \( N_{WG} \) is the total number of wind units. Also the input power to dc/ac convertor can be calculated as:

\[
P_i^L(t) = \frac{P_i^{Load}(t)}{n_i}
\]

(15)

In which \( P_i^L(t) \) is the input power to convertor, \( n_i \) is the convertor efficiency and \( P_i^{Load}(t) \) is the estimated load demand in \( i \)th interval and \( I_{th} \) day. Battery capacity can be calculated with this process:

*If \( P_i^{ref}(t)=P_i^L(t) \) then battery capacity would not change.*

*If \( P_i^{ref}(t)>P_i^L(t) \), then additional power amount which is \( P_i^{ref}(t)-P_i^L(t) \) is used for battery charging.*

*If \( P_i^{ref}(t)<P_i^L(t) \), then the lack of power or \( P_i^{ref}(t)-P_i^L(t) \) can be compensated with the battery itself.*

All above-mentioned equations and the resulting model with respect to its constraints run in hourly and yearly running intervals. The flowchart of our algorithms is show in Fig. 4.

### 2.5. GSA Algorithm

GSA is an algorithm which is related to collective intelligence and of course it is memory less. This
optimization algorithm is designed based on gravitational rules and object movements in an artificial system at discrete times in which system space is the problem definition space. With respect to gravity rule, any object (with mass) understands the location and position of other objects by gravitational rule. So we can use this force as a tool for information exchange [10]. In this algorithm object masses are determined with respect to objective function. In a system with n objects, the position of each object is a point in the space that is a solution for our problem. $X_i$ is the position of the object in dth dimension and can be shown as:

$$X_i = (x_i^1, x_i^2, ..., x_i^n) \text{ for } i = 1, 2, ..., N$$  \hspace{1cm} (16)

In which, $n$ is the dimension of the problem and $N$ is the number of objects. In this system a force with the amount of $F_{ij}^d(t)$ is applied to $I_i$ object from $J_{th}$ object at the time of $t$ and in the direction of d. The amount of this force can be formulated as:

$$F_{ij}^d(t) = G(t) \frac{M_{pi}(t) \times M_{pj}(t)}{R_{ij}^d(t)^2} (x_i^d(t) - x_j^d(t))$$  \hspace{1cm} (17)

In which $M_{pi}(t)$ is the active gravitational mass of j and $M_{pj}(t)$ is the inactive gravitational mass of I, and in our algorithm we assume that both of them are equal to $M$. Also each object has an acceleration and speed which are respectively shown by $a_i(t)$ and $V_i(t)$. With respect to the second law of Newton each object in the $d_{th}$ dimension has an acceleration which is proportional to the applied force (in that dimension) divided by inertial mass and this is shown in (18). Moreover the speed of each object in time can be calculated by (19):

$$a_i^d(t) = \frac{F_{ij}^d(t)}{M_i(t)}$$  \hspace{1cm} (18)

$$v_i^d(t+1) = rand_j \times v_i^d(t) + a_i^d(t)$$  \hspace{1cm} (19)

After the calculation of speed and acceleration, we can use (20) in order to calculate the new position of $I_{th}$ object in $d_{th}$ dimension.

$$x_i^d(t+1) = x_i^d(t) + v_i^d(t+1)$$  \hspace{1cm} (20)

In our method, new position is considered as the location of new objects in search space and we can normalize the mass of these new objects by (21) and (22):

$$m_i(t) = \frac{fit_i(t) - worst(t)}{best(t) - Worst(t)}$$  \hspace{1cm} (21)

$$M_i(t) = \frac{m_i(t)}{\sum_{j=1}^{n} m_j(t)}$$  \hspace{1cm} (22)

In which $fit_i(t)$ represents the amount of fitness of the $I_{th}$ object at t and worst(t) and best(t) are respectively the fitness amounts of the worst and the best objects in the total population at t and we can calculate their amounts by these equations:

$$best(t) = \min \{ fit_i(t) \}$$  \hspace{1cm} (23)

$$worst(t) = \max \{ fit_i(t) \}$$  \hspace{1cm} (24)

2.6. PSO Algorithm

PSO algorithm is inspired from nature and is based on social behaviour of birds of fishes in the food finding processes. Movement rules and searching rules in this algorithm are simple and meaningful and are developed for NLP programs with continues variables but also we can use them for NLP problems with discreet variables too. In this algorithm the location of each particle changes by its speed vector. Direction and amount of each speed vector is also influenced by previous speed vector in the direction of the best previous personal and group experience. The mathematical representation of this thing is shown in eq. (25).

![Fig 4. Flowchart of the proposed optimization algorithm for the simulated model](image-url)
As we can see, the previous speed vector is combined with distance vector to the best personal result and distance vector to the best group result and so the new speed vector direction is determined:

\[ \mathbf{V}_i^{k+1} = w \times \mathbf{V}_i^k + C_1 \times \text{rand}_1 \times (\mathbf{P}_\text{Best}_i - \mathbf{S}_i^k) + C_2 \times \text{rand}_2 \times (\mathbf{g}_\text{Best} - \mathbf{S}_i^k) \quad (25) \]

In which \( \mathbf{V}_i^{k+1} \) is the corrected speed vector for \( i_{th} \) particle in \( K+1 \)th iteration, \( \mathbf{V}_i^k \) is the speed vector for \( i_{th} \) particle in \( K \)th iteration, \( \mathbf{S}_i^k \) is the location coordinate of \( i_{th} \) particle in \( k \)th iteration, \( \text{rand}_{1,2} \) is a random number between 0 and 1, \( \mathbf{P}_\text{Best}_i \) is the location vector of the best personal experience (for \( i_{th} \) particle), \( \mathbf{g}_\text{Best} \) is the location vector of the best group experience, \( w \) is the weight factor for each particle, and \( C_1 \) and \( C_2 \) are learning coefficients which are fixed on 2. With eq. (25) for each particle we can calculate a special speed and in the next iteration we can rewrite particle location as:

\[ \mathbf{S}_i^{k+1} = \mathbf{S}_i^k + \mathbf{V}_i^{k+1} \quad (26) \]

\( w \) or weight function can be determined by eq. (27). Regularly the amount of this weight factor in each iteration, decreases linearly so it can be guaranteed that the direction is toward the best personal and group experience:

\[ w = w_{\text{max}} \times \left(1 - \frac{\text{iter}}{\text{iter}_{\text{max}}}\right) \quad (27) \]

In which \( w_{\text{max}} \) is the initial weight, \( w_{\text{min}} \) is the final weight; \( \text{iter}_{\text{max}} \) is the maximum allowable iteration and \( \text{iter} \) is the number of iteration.

3. Results and Discussion

3.1. Test System 1

In this test system, the proposed model have been simulated in MATLAB 7.1 and IEEE-30 bus reliability test system (IEEE-RTS30) with 1 kW maximum peak load or load demand has been used as a load profile. Other assumptions like PV modules installation tilt angle and simplification in simulation time-intervals are obtained from [2]. The average daily vertical and horizontal solar irradiation during 52 weeks of a year is shown in Fig. 5 (left). Also, Wind speed at the WG installation height is shown in Fig. 5 (right). In addition, information of load profile is presented in Fig. 6 (left). Simulation of the proposed model was lasted 5 seconds.

3.2. Test system 2

In this test system, the proposed model have been simulated in MATLAB 7.1 and IEEE-30 bus reliability test system (IEEE-RTS30) with 10 kW maximum peak load or load demand has been used as a load profile.

<table>
<thead>
<tr>
<th>System Type</th>
<th>Number of PV modules</th>
<th>Number of WG</th>
<th>Number of Batteries</th>
<th>Total Cost (Euro)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid PV/WG</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>26129.09</td>
</tr>
</tbody>
</table>

Fig 5. 24-hour average irradiation during 52 weeks of a year (left figure) and 24-hour average wind speed during 52 weeks of a year (right figure)

The algorithm results or optimum combination is presented in Table 1. It should be mentioned that information about the types of equipment and their costs is obtained from [1] and [2]. Due to the continuous essence of GSA and the discontinuity of this problem results, value of objective function is determining as the nearest value to the discontinuous result of each variable in each iteration. As a matter of fact, answers will be rounded. Fig. 6 (right) shows the battery charge of the optimum combination during the case study. Also, Fig. 7 shows the convergence procedure of GSA in solving duration for PV/WG sizing problem.
Other assumptions like PV modules installation tilt angle and simplification in simulation time-intervals are similar to the test system 1 [2]. The average daily vertical and horizontal solar irradiation during 52 weeks of a year and Wind speed at the WG installation height are similar to the test system 1 too. In addition, information of load profile is presented in Fig. 8 (left). Simulation of the proposed model was lasted 112 seconds. The algorithm results or optimum combination is presented in Table 2. Fig. 8 (right) shows the convergence procedure of GSA for PV/WG sizing problem. Fig. 9 shows the convergence procedure of PSO algorithm for PV/WG sizing problem.

**Fig 6.** 24-hour load demand during 52 weeks of a year of the test system 1 (left) and the battery charge of the optimum combination during the case study (right).

**Fig 7.** Convergence procedure of the proposed algorithm

**Table 2.** Optimum combination of the hybrid PV/WG system according to the proposed model

<table>
<thead>
<tr>
<th>System Type</th>
<th>Number of PV Modules</th>
<th>Number of WG</th>
<th>Number of Batteries</th>
<th>Total Cost (Euro)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid PV/WG System</td>
<td>2</td>
<td>30</td>
<td>41</td>
<td>214356.85</td>
</tr>
</tbody>
</table>

**Fig. 8.** 24-hour load demand during 52 weeks of a year of the test system 2 (left) and Convergence procedure of the GSA algorithm in test system 2 (right)
4. Conclusion

Application of renewable energy sources as a distributed generation will reduce the emission of greenhouse gases and also, will save a lot of investment that lost in fossil energy fields. There are a lot of potentials to use renewable energy sources in Iran. This paper presents the optimum sizing of a hybrid PV/WG system for three test systems, using gravitational search algorithm (GSA). The results show the algorithm ability to solve optimization problems in power systems in a fast and accurate manner and also, without any computational burden.

References


CFMTL: Clustering Wireless Sensor Network Using Fuzzy Logic and Mobile Sink In Three-Level

Ali Abdi Seyedkolaei1 and Ali Zakerolhosseini2

1 Department of Computer, Shahid Beheshti University, Tehran, Iran
   Aabdi64@gmail.com

2 Department of Computer, Shahid Beheshti University, Tehran, Iran
   Ali.zaker@sbu.ac.ir

Abstract
Lifetime enhancement has always been of crucial importance for energy constrained sensor network due to resource limitations of sensor nodes. The protocol play important roll, which can minimize the delay while offering high energy efficiency and long span of network lifetime. This paper concentrate on energy optimization by introducing a novel and an adaptive clustering algorithm that is fuzzy logic and mobile sink based. In this paper, a three-level fuzzy logic is utilized to evaluate the Priority of sensors to become a cluster head. In the first level, the qualified nodes are selected based on their residual-energy and density of them. Then, in the second level, nodes overall cooperation is considered in the whole network with three fuzzy parameters. These parameters are centrality, proximity to base station and distance be- tween cluster-heads. In the third level, the sink is move based on two parameters energy of cluster heads and distance of cluster heads to sink. This paper compares the result of this CFMTL approach to LEACH, Fuzzy, GA and FGA algorithms.

Keywords: Wireless Sensors Network, Network lifetime, Energy Efficient, Fuzzy logic, Mobile Sink.

1. Introduction
Advances in Micro Electro-Mechanical System (MEMS) technologies, embedded computing technologies and wireless communication technologies have enabled the development of relatively inexpensive and low-power-consumption micro sensors with the capability of sensing, computing and communicating. Composed of a large number of these sensor nodes, a Wireless Sensor Network (WSN) can be used for detecting, collecting and analyzing the information of complex environments in real time. WSNs can be used in many applications such as military, biomedical, and environmental applications. WSN routing technology is one of the most important technologies. According to the network architecture, WSN routing technologies can be divided into planar and hierarchical, and hierarchical network is also referred to cluster based. Many experiments have provided evidence that cluster based routing protocols work better in network topology management, energy minimization, data aggregation and so on than planar ones [1]. In large scale networks, nodes are usually partitioned into a number of small groups, called clusters, for aggregating data through efficient network organization. In cluster based networks, cluster head nodes collect data packets from cluster members and transmit data to base station (BS) in either single hop or multi-hop mode by means of wireless communication.

Utilizing intelligent techniques improves the efficiency of wireless sensor network. In applications that require real-time decision making, fuzzy logic is a powerful tool that can make decision even if there is insufficient data; while sufficient data (which is rare in real applications) is needed for making a decision in classic control. Recently, in some papers like [2] and [3], fuzzy logic is used for routing and improving network lifetime. We also used fuzzy logic as a mean to select cluster heads. In addition, due to the proved efficiency of clustering techniques in energy consumption, we utilized clustering in our proposed routing algorithm.

In this paper, we used a fuzzy logic in three-level to evaluate the Priority of sensors to become a cluster head and also transmission sink to suitable place to reduce energy consumption. In first level the qualified nodes are selected based on their residual-energy and density of them. Then, in the second level we seek for the best node cooperation regarding to the average energy consumption metric. In third level the sink is move based on two parameters energy of cluster heads and distance of cluster heads to sink.

In the remaining of this paper, Section 2 discusses some related works and previous studies. Section 3 describes the system and energy model and Section 4 discusses proposed algorithm. Section 5 provides simulation results and discusses the efficiency of proposed algorithm. Finally, Section 6 gives concluding remarks.
2. Related Work

In this section, we review related work in clustering algorithms. A main issue in the design of wireless sensor networks is the power dissipation scheme, hence the wireless node has a limited energy tag battery and has no backup power source until node death, thus, researches consider the design of low-power signal processing architectures, low power sensing interfaces, energy efficient wireless media access control and routing protocols, which revolves around energy balancing and management process.

LEACH [4] is one of the first hierarchical routing approaches for sensors networks, which attempts to improve energy and routing efficiency of such networks. The idea proposed in LEACH has been an inspiration for many hierarchical routing protocols, although some protocols have been independently developed. In HEED protocol, residual energy of each sensor node is the primary parameter for probabilistic election of cluster-heads [5]. In case of a tie in cluster-head election, node degree or average distance to neighbors parameters are used to determine the cluster-head. Experimentations that are employed for evaluating HEED protocol show that clustering and data aggregation at least double the lifetime of the WSN.

In [6] authors have presented Intelligent Fuzzy-based cluster head selection system for WSNs and the performance has been analyzed. Selection of cluster head is difficult in different situations having different characteristics. Based on fuzzy theory and number of neighbor nodes, an energy efficient algorithm known as F3N has been developed by researchers. They have presented F3N algorithm for cluster head election using fuzzy theory. In [2], Gupta used fuzzy logic to find cluster heads. In this algorithm three fuzzy variables is used for cluster head selection. Nodes energy, nodes concentration and nodes centrality are these parameters. In this approach, the base station primarily collects the necessary information from all nodes and then selects a node as a cluster head according to the fuzzy rules. In this approach there is only one selected CH for each round, whereas more CHs are needed for balancing energy consumption and improving network lifetime.

In [3], Kim offers CHEF in which, the same as [2], the CHs are selected based on a fuzzy logic. The difference is that in this approach more than one cluster head is selected locally in each round. The fuzzy set includes nodes energy and their local distances. CHEF [3] also generates a random number for each sensor and if it is less than a predefined threshold, Popt, then the nodes chance is determined. Thus, there may be some qualified nodes that lose their chance on a random manner.

In this paper, we used fuzzy logic in three-level to determine CHs in each round and also transmission sink to suitable place for prolonging network lifetime to an acceptable limit.

3. System and Energy Model

The mentioned network has following characteristics:
1. Nodes are randomly spread in the environment and nodes have been assumed homogeneous.
2. Initially, Base Station is located in the center of the environment.
3. Nodes are able to adjust their sending power according to their distance to the intended receiver.
4. All nodes have equal energy and ability.
5. Location and ID for all nodes is known for base station.
6. Sink is mobile and is moving on a square path.

Energy consumption model used in this article, is the same as energy consumption model in the LEACH [4] article. Each node to send 1 bit data to d distance of itself consumes as much as $E_s$ energy, which is obtained from Eq. (1):

$$E_s = \begin{cases} IE_{select} + IE_{f}d^2 & d < d_{co} \\ IE_{select} + IE_{f}d^4 & d \geq d_{co} \end{cases}$$

(1)

Also the amount of energy that is used in the receiver for receiving 1 bit node is obtained from Eq. (2):

$$E_r = IE_{select}$$

(2)

The assumption is that in each period, a cluster head receives only a packet from each node of its cluster. After receiving all packets, cluster head reports their useful information in the form of a single packet in a multi-hop manner to base station.

4. Proposed Algorithm

In this section, our proposed algorithm are presented to increase the network lifetime and energy consumption. Our method is composed of three levels that in the first level based on two parameters of the residual-energy and density is denoted cluster heads candidate each region. Density is the number of nodes adjacent to a node that will be useful range. Selecting region is that each node radius of R, all nodes within the radius is considered as the neighborhood and constitutes a region. In this level selection cluster
heads will be determined as Distributed. The cluster heads which are selected, are not final and only for best selection overall are transmitted to next level. The membership functions of these parameters are depicted in Fig. 1, 2 and 3. The fuzzy if-then rules in the first level are also shown in Table 1.

In second level, cluster heads selection will be based on three parameters, centrality, proximity to base station and distribution between cluster heads that is their selection as focus. Centrality parameter defines that cluster head to what extent is in the center of a cluster. Whatever of this amount be less than, expresses is the fact that normal nodes are located at a distance less than cluster head and this causes reduce the energy consumption of normal nodes during data transfer to its cluster head. The second parameter (proximity to base station) causes that sending data to base station reduced if it is close to cluster head. and the third parameter means distribution between cluster heads, in case of high its value, reduces energy consumption because whatever is much more, suggest that distance between cluster heads increased and better distribution have in whole network. If this value be lower, Causes aggregation cluster heads in area high and reduced distances to other areas and if normal nodes want send data, they consume much energy. In this case, the network lifetime reduces before the right time and the first dead node appears. The membership functions and the fuzzy rules of these parameters are depicted in Figures 4 to 7 and Table 2 respectively.

<table>
<thead>
<tr>
<th>Residual-Energy</th>
<th>Density</th>
<th>Priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Low</td>
<td>Very Small</td>
</tr>
<tr>
<td>Low</td>
<td>Medium</td>
<td>Small</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>Rather Small</td>
</tr>
<tr>
<td>Medium</td>
<td>Low</td>
<td>Medium Small</td>
</tr>
<tr>
<td>Medium</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Medium</td>
<td>High</td>
<td>Medium Large</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>Rather Large</td>
</tr>
<tr>
<td>High</td>
<td>Medium</td>
<td>Large</td>
</tr>
<tr>
<td>High</td>
<td>High</td>
<td>Very Large</td>
</tr>
</tbody>
</table>
In the third level, we prioritize the cluster head based on two parameters, energy and distance to the sink through fuzzy logic. We consider cluster heads with low energy and long distance as the cluster head high priority and move the sink on the considered square path to be much close to a cluster head which is placed with high priority. We perform this operation in each round between the cluster heads. This operation causes the reduction of energy consumption and the increase of network lifetime. The membership functions and the fuzzy rules of these parameters are depicted in Figures 8 to 10 and Table 3 respectively.

In this algorithm single-hop inter cluster and multi-hop intra cluster communications are used there on of using multi-hop communication is due to the Cluster heads that has further communication with the base station after collecting data from its cluster nodes which will send it to the nearest cluster head until will consume less energy to send their data directly to the base station and data with multi-hop intra cluster communication in each round are

<table>
<thead>
<tr>
<th>Centrality</th>
<th>Proximity to BS</th>
<th>Distance between CHs</th>
<th>Priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Low</td>
<td>Medium</td>
<td>Large</td>
</tr>
<tr>
<td>Low</td>
<td>Low</td>
<td>High</td>
<td>Very Large</td>
</tr>
<tr>
<td>Low</td>
<td>Medium</td>
<td>Low</td>
<td>Rather Large</td>
</tr>
<tr>
<td>Low</td>
<td>Medium</td>
<td>Medium</td>
<td>Large</td>
</tr>
<tr>
<td>Low</td>
<td>Medium</td>
<td>High</td>
<td>Large</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>Medium Large</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>Medium</td>
<td>Rather Large</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Rather Large</td>
</tr>
<tr>
<td>Medium</td>
<td>Low</td>
<td>Low</td>
<td>Medium</td>
</tr>
<tr>
<td>Medium</td>
<td>Low</td>
<td>Medium</td>
<td>Medium Large</td>
</tr>
<tr>
<td>Medium</td>
<td>Low</td>
<td>High</td>
<td>Medium Large</td>
</tr>
<tr>
<td>Medium</td>
<td>Medium</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Medium</td>
<td>High</td>
<td>Low</td>
<td>Rather Small</td>
</tr>
<tr>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td>Medium Small</td>
</tr>
<tr>
<td>Medium</td>
<td>High</td>
<td>High</td>
<td>Medium Small</td>
</tr>
<tr>
<td>Medium</td>
<td>High</td>
<td>High</td>
<td>Medium Small</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>Small</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>Medium</td>
<td>Rather Small</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>Rather Small</td>
</tr>
<tr>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>Very Small</td>
</tr>
<tr>
<td>High</td>
<td>Medium</td>
<td>Small</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Medium</td>
<td>High</td>
<td>Small</td>
</tr>
<tr>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Very Small</td>
</tr>
<tr>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Small</td>
</tr>
<tr>
<td>Energy</td>
<td>Distance</td>
<td>Priority</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>Near</td>
<td>Very Small</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>Med</td>
<td>Small</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>Far</td>
<td>Rather Small</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>Near</td>
<td>Medium Small</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>Med</td>
<td>Medium</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>Far</td>
<td>Medium Large</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Near</td>
<td>Rather Large</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Med</td>
<td>Large</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Far</td>
<td>Very Large</td>
<td></td>
</tr>
</tbody>
</table>

Copyright (c) 2014 Advances in Computer Science: an International Journal. All Rights Reserved.
sent to the base station. The way to choose a nearest cluster head which each cluster head of radius $R$, selects the head clusters according to proximity.
The important point is that in this algorithm in case a node proximity the base station to its cluster head can send their data directly to the base station which reduces the energy consumption in the network and this makes the method priority to the other methods.
Clustering operation is performed in each period, but it does not have to be changed in each step of the previous cluster head node because some Parameters which were considered for the fuzzy module, may distinct it among the other nodes. The Following fig. 11 shows the structure of the network after complete cover the network. The red node show cluster-head and green ones are those which are connected to cluster-head and blue line show route send data from cluster head to sink.

5. Simulation and Results

We have simulated the proposed algorithm and other protocols with the MAT- LAB software. We will compare the CFMTL algorithm with LEACH [4], Fuzzy [7], GA [8] and FGA [9]. The parameters considered in this simulation are given in Table 4.

5.1 First-dead Node (FDN) Compare

Table 5 displays the dead time of the first node in the proposed algorithm and other methods in different environments.

<table>
<thead>
<tr>
<th>Algorithm Type</th>
<th>Network Parameters</th>
<th>LEACH</th>
<th>Fuzzy</th>
<th>GA</th>
<th>FGA</th>
<th>CFMTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>150 node with network size : 250*250</td>
<td>78</td>
<td>92</td>
<td>118</td>
<td>150</td>
<td>338</td>
<td></td>
</tr>
<tr>
<td>150 node with network size : 250*250</td>
<td>15</td>
<td>17</td>
<td>42</td>
<td>95</td>
<td>253</td>
<td></td>
</tr>
<tr>
<td>200 node with network size : 400*400</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.2 Consume Energy

In figure 12, we have compared the proposed algorithm with other methods in terms of reduction rate of the networks energy. As can be clearly seen, our algorithm has made the reduction rate of the networks energy much smoother and more uniform compared with other methods.

5.3 Number of Alive Nodes

Improving performance of the algorithm in terms of increased network lifetime compared with other methods is shown in figure 13. As can be seen, the proposed algorithm has increased the death time of the first node in comparison with other methods. Even after the death of the first node of the network, if the network will be allowed to proceed until all of the networks nodes consume their
energy, we can observe that even after the death of the last node, the number of the periods spent would still be higher.

Fig. 13 Compare the number of live nodes in the network size: 400*400, number of nodes: 200

6. Conclusion

Optimum energy consumption in wireless sensor networks is of great importance, so that optimum consumption of energy leads to increasing networks lifetime. An efficient routing technique is known as hierarchical routing based on clustering that prolongs the network lifetime. In this paper, the most priority cluster heads were selected via a three-level fuzzy logic. In the first level the priority nodes were selected based on their residual-energy and density. Then, in the second level, nodes overall cooperation is considered according to centrality, proximity to BS and distance between CHs in the whole network. In the third level, the sink is move based on two parameters energy of cluster heads and distance of cluster heads to sink. The proposed algorithm was compared with similar approaches LEACH [4], Fuzzy [7], GA [8] and FGA [9] in energy consumption and number of live nodes. The performance of the algorithm was evaluated by a simulation and the results showed that in this approach, nodes consume less energy and live longer. Moreover, a fair load distribution and hence fewer variance of energy consumption demonstrate the efficiency of the proposed algorithm.

References


Towards a Secure Maturity Model for Protecting e-Government Services in Tanzania: A Stakeholders View

Mohamed D. Waziri¹ and Zaipuna O. Yonah²
The Nelson Mandela African Institution of Science and Technology (NM-AIST)
School of Computational and Communication Science and Engineering
P. O. Box 447, Arusha, Tanzania
Email: {dewam¹, zaipuna.yonah²}@nm-aist.ac.tz

Abstract

E-Government Maturity Models (eGMMs) are widely used as a tool for guiding the development and implementation of e-Government services. The government of Tanzania recognizes that e-Government services can accelerate the achievement of a sustainable social and economic development in the country. However, despite the good benefits provided by e-Government services, information security and privacy are the most significant obstacles for e-Government services adoption. Unfortunately, very few designs of e-GMMs have considered security as a specific issue. However, even these few security responsive models consider security mostly at the transaction stage. Responding to this security weakness of eGMMs, in our earlier work a holistic secure e-Government maturity model that includes security layers consisting of technical and non-technical security related aspects in each of its four maturity stages was developed, but the model was not yet tested and evaluated. This paper reports on the tests and the validation of the proposed secure model. The applied evaluation criteria were: simplicity, reliability, accessibility and usability, dynamics and flexibility, applicability, coverage and completeness, and compliance with legal aspects. Primary data were collected from five Tanzanian public organization using questionnaires. The collected data were processed and analyzed using the SPSS. The overall results show that the model designs meet all required specifications to successfully secure e-Government services, and the model was accepted by majority of the respondents at different organizational levels (strategic, tactical, and operational).

Keywords: e-Government, e-Government services, e-Government maturity model, information security, security layer

1. Introduction

With the recent advancements of Internet technologies, majority of governments around the world have adopted Information and Communication Technologies (ICTs) to provide services towards its agencies, businesses and citizens more efficiently and effectively. In general, e-Government refers to the use of ICTs by government organizations to provide and enhance delivery of public services. In the Tanzanian context, e-Government is about delivering quality services to the public through technology [1]. It involves using ICTs to support processes within the government as well as for the delivery of services to beneficiaries, such as citizens, businesses and agencies. In practice, e-Government websites or portals provide governments the best opportunities to improve their administrative processes and procedures, to connect to their citizens effectively as well as to build and bond interactively with its agencies and businesses [2].

Tanzania is one of many developing countries with multiple e-Government initiatives being introduced to support poverty reduction, and sustain good governance, as demonstrated by recent technology implementations and reported in government strategy documents [3]. The government of Tanzania recognizes that through e-Government services a sustainable social and economic development in the country is achievable. As a result, the Government has increased the range and quality of the services provided by public sector [1]. The Government also recognizes the importance of e-Government in promoting and improving efficiency in public services delivery and strengthening citizen’s participation and engagement. However, despite the good benefits provided by e-Government services, a number of obstacles exist that hinder achieving these desired benefits. Specifically, information security and privacy are among the most significant obstacles faced when implementing e-Government initiatives [4]. As a result, there are increasing concerns about the reliability and security of the developed e-Government services. These concerns have flagged the need to guarantee and to ensure that services are provided to customers with the maximum possible security, with guaranteed privacy [5]. The current trend is that citizens prefer to use traditional ways rather than using an unsecured e-Government service. Noted also is that citizens’ adoption of e-Government services plays an important role in the success of e-Government initiatives. Thus, low adoption, particularly by citizens, indicates inadequate utilization and rejection of the initiatives by the intended users, and this may lead into failure of e-Government initiatives [6].

Various researchers have proposed different methods and systems to provide security in e-
Government services. In [7], Wimmer and Bredow propose a comprehensive way to holistically approach that integrates security aspects from the strategic level down to the data and information level in order to address different security aspects of e-Government services. Their holistic approach consists of four layers, namely: strategic, process level, interaction and information.

In the Tanzanian context reported in [8], Dewa and Yonah propose a holistic secured e-Government Maturity Model consisting of both technical and non-technical security aspects for protecting e-Government services with Tanzania as a case study. With a holistic approach, security is considered beyond the technical aspects. Social, political, cultural, and legal impacts on security requirements are considered as well [9]. The model design process was based on ISO/ IEC 27002 and was guided by a Design Science Research methodology (DSR). The applied DSR steps were: problem identification and motivation, definition of the objectives for a solution, and development of the model [10]. However, one step remained to be applied that is: model evaluation, to test if the model meets the specifications and that it fulfils its intended purpose. Thus, the purpose of this paper is to report on the tests and validation of the holistic secure maturity model for protecting e-Government services with Tanzania as a case study proposed in [8].

The rest of the paper is organized as follows: Section two presents the background of the designed secure maturity model; Section three outlines the research methodology; Section four presents the results and discussion; Section five outlines recommendations; and lastly, the conclusion is given in Section six.

2. Background

Information security and privacy have been widely recognized as the main obstacles to the adoption of e-Government services. In [11], Dewa and Zlotinikova identify the information security requirements to e-Government services. Those security requirements include confidentiality, integrity, availability, non-repudiation, authentication, authorization (access control), traceability, accountability, user anonymity, and security awareness. However, e-Government services may not require the application of all identified security requirements. Typically, each e-Government service has its own specific security requirements depending on the services it provides. Practically, information security of e-Government services is influenced by how these services were developed, and e-Government Maturity Models (eGMMs) are widely used as a tool for guiding the development and implementation of e-Government services. An eGMM is a set of stages (from basic to advanced ones) that determines the maturity of the e-Government [12]. For instance, West proposes a three stage model [13], Layne and Lee propose a four stage model [14], Hiller and Belanger propose a model with five stages [15], and Deloitte and Touche propose a six stage model [16]. The critical weakness of the existing models is the consideration of security related issues at the transaction stage only [17]. In order to secure eGMMs, both technical and non-technical security related aspects should be considered at all stages of the maturity models. Responding to the eGMMs weakness, a secure e-Government maturity model was developed as presented in [8]. The following paragraphs briefly describe the model.

A secure e-Government maturity model includes security layers that consist of technical and non-technical security related aspects at each of the four proposed critical stages of the model. The critical stages of the model are: (a) secured digital presence, (b) secured interaction, (c) secured transaction, and (d) secured transformation [8]. It was recommended that implementation of the model should be based neither on a specific technology/protocol nor on a certain security system/product, but rather be based on an approach towards a structured and efficient implementation of those technologies.

At secured digital presence stage, the security layer should have the ability to verify e-Government services identity in order to build trust between government agencies and users. The most important security related aspects to be considered at this stage are information availability and entity authentication. The security controls at this stage aim at preventing unauthorized physical access or interference with the organization or ICT equipment and information assets.

The second stage of the model is secured interaction stage. At this stage, security layer should have the ability to authenticate a user/citizen asking for a service. The most important security aspects of this stage are identity authentication, availability and integrity. These aspects can be achieved through the implementation of all security practices required at secured digital presence stage together with the implementation of database security controls, audit management and the presence of the adequate bandwidth capacity.

At secured transaction stage the most important security aspects are personal information confidentiality, identity authentication, availability, non-repudiation, accountability and integrity. At this stage, the security layer should include the implementation of certificate/digital signature and secure data transmission in order to achieve data
integrity and confidentiality of citizens’ personal information. The exchanged message should be encrypted in order to ensure their confidentiality. The data contained in the e-Government services and exchanged between the different government agencies must remain confidential.

A security layer at secured transformation stage should restrict the utilization of personal information, and secure such information from access by unintended parties. A government agency should be able to authenticate another government agency that requires a service on behalf of the users. The security layer should also have the capability of filtering service access, because some agencies will not have the right to invoke a certain service while others do. The most important security aspects of this stage are personal information confidentiality, identity authentication, availability, non-repudiation, accountability and integrity. At this stage, access control mechanisms should be implemented together with other security practices.

2.1 Objectives

The general objective of the study reported in this paper was to test and validate the proposed holistic secure maturity model for protecting e-Government services with Tanzania as a case study. Specific objectives of this study were as follows:

a. To identify the criteria used to evaluate the model.

b. To test and validate the model.

c. To recommend activities for improving the model and security of e-Government services in Tanzania.

3. Methodology

For the purpose of testing and validating the proposed holistic secure maturity model for protecting e-Government services in Tanzania, data were collected using structured questionnaires. The questionnaires were distributed to staffs with ICTs skills or ICTs security expertise at three different organizational levels: strategic, tactical, and operational. In order to achieve a comprehensive model evaluation, it was important to identify evaluation criteria. Based on literature [18-21] seven model evaluation criteria were selected. The identified criteria were: simplicity, reliability, accessibility and usability, dynamics and flexibility, applicability, coverage and completeness, and compliance with legal aspects. The validation plan is further explained in the following sub-sections.

3.1 Population and Sampling Method

A population of government organizations staff was consulted to test and validate the model. Specifically staffs with ICTs skills or ICTs security expertise at strategic, tactical, and operational levels were consulted. The targeted research sample involved the five government organizations studied earlier [11]. Due to confidentiality reasons, we referred them as Organizations A, B, C, D and E as follows [11]: Organization A is a public organization responsible for managing the overall revenue, expenditure and financing of the government. Organization B is a public organization responsible for management of public services. Organization C is a public organization responsible for managing the assessment, collection and accounting of all central government revenue. Organization D is a public organization responsible for generating, transmitting, distributing and selling electricity. Organization E is a public organization responsible for coordinating, encouraging, promoting and facilitating investment. Initially, the sample size of the population was estimated at 60. We distributed 70 questionnaires, and 45 responses were received.

Table 1: Summary of the respondents in the studied organization

<table>
<thead>
<tr>
<th>Organization name</th>
<th>Strategic Level</th>
<th>Tactical Level</th>
<th>Operational Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>E</td>
<td>-</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>8</td>
<td>33</td>
</tr>
</tbody>
</table>

3.2 Data Collection

Primary data were collected by using the questionnaires. In order to ensure validity of the questionnaires, a pilot study was conducted prior to distributing the questionnaires to respondents. Twenty five questionnaires were delivered to the respondents, but only 11 responded. Necessary improvements to the questionnaire were done, and the improved version of the questionnaires was distributed to the sample population. Table 1 shows the distribution of the respondents within the studied organizations.

3.3 Reliability and Goodness of Fit Measurement.

Analysis of internal consistency of the questions was conducted by doing a reliability test by using the Statistical Package for Social Sciences (SPSS). The calculated Cronbach’s alpha was 0.966 and
since it is generally agreed that an alpha of 0.70 and above is acceptable [22], therefore, the calculated Cronbach’s alpha was found suitable. To assess goodness of fit for all survey questions, a Chi-square test was conducted and the results showed statistical significance because all the variables had p-value 0.000 (degree of freedom = 4) which is less than any level of significance that is 1%, 5% and 10%.

3.4 Data Processing and Analysis

The content analysis technique was used for processing and analyzing descriptive data. SPSS and Microsoft Excel were used for data analysis. Various analysis techniques including descriptive statistics, especially frequency and a Chi-square test were used to measure both goodness of fit and relationship between variables. The results were presented using charts as presented in Section 4.

4. Results and Discussions

4.1 Model Testing and Validation

The purpose of this paper was to test and validate the model. The aim of this process is to evaluate the proposed model to assess if it is correct, complete, being implemented as intended, and delivering the intended outcome. Typically, an evaluation process involves comparing the objectives of a solution to actual observed results from the use of the developed model. In our case, this was done through identification and analysis of the model evaluation criteria. The identified criteria were as follows: (1) simplicity; (2) reliability; (3) accessibility and usability; (4) flexibility; (5) applicability; (6) coverage and completeness; and (7) compliance with legal aspects. As mentioned earlier, we received 45 responses, and these responses were processed and analyzed by using the Statistical Package for Social Sciences.

4.1.1 Simplicity

Simplicity is the state or quality of being easy to understand or use. The aim of this evaluation criterion was not only to assess if the model was designed in such a way that it is clear and easily understandable to e-Government services implementation team, but also to assess if the model was designed in such a way that it is easy to implement. Figure 1 shows the results of the analysis of collected data against the simplicity criterion of the model clarity and comprehensibility.

![Figure 1](image)

Respondents’ responses on the model clarity and comprehensibility satisfaction.

Evident from this figure is that 82.2% (summation of 71.1% and 11.1%) of the respondents were satisfied with the simplicity of the model in terms of its clarity and comprehensibility to e-Government services developers. The analysis also shows that 75.6% (summation of 66.7% and 8.9%) of the respondents were also satisfied with the simplicity of the model in terms of the model implementation as shown in Fig. 2. These results imply that the respondents observed the model design to be clear and easily understandable to e-Government developers and implementers. Therefore, the model design reflects and considers the implementation environment in terms of in-house e-Government services developers’ skills.

![Figure 2](image)

Respondents’ responses on the model ease to implement satisfaction.

4.1.2 Reliability

Reliability is a quality of being reliable, dependable or trustworthy. The aim of this evaluation criterion was to assess if the model performs its required functions under stated conditions for a specified...
period of time. Specifically, this criterion assesses if the implementation of the model is capable of treating security risks and threats posed to the current e-Government services consistently. Information security risk assessment provides organizations a capability of discovering, correcting and preventing security problems in e-Government services. The risk assessment helps each organization to determine the acceptable level of risk and the resulting security requirements for each system.

The results show that 84.4% (summation of 71.1% and 13.3%) of the respondents were satisfied with the reliability of the model. This implies that the model design considers information security risk assessment and treatment consistently. Figure 3 shows the respondents’ responses on the model reliability satisfaction.

The results show that 66.6% (summation of 62.2% and 4.4%) of the respondents were satisfied with the accessibility and usability of the model. In addition, the results show that the model designs have the capacity to make sure that security aspects do not compromise accessibility and usability of e-Government services.

### 4.1.4 Flexibility

Flexibility is the ability to be easily modified or responsive to change. Given that ICTs are advancing rapidly, computer hackers and attackers are also advancing their techniques. Attackers are always coming up with new ways to defeat the improved security protection. In this situation, any security design tool or method should have a capability of being flexible to catch up any technological advances. Therefore, the aim of flexibility criterion was to assess if the model is dynamic enough to deal with possible future security risks and threats. Figure 5 shows respondents’ responses on the model flexibility satisfaction.

The results show that 71.1% (summation of 64.4% and 6.7%) of the respondents were satisfied with the capability of the model to respond rapidly with the technology advancement to deal with possible future security risks and threats. This outcome implies that the model design enables the organization to review security risks and threats of the e-Government services periodically.
4.1.5 Applicability

Applicability is the state of being pertinent, relevant or appropriate. The aim of this evaluation criterion was to assess if the model was designed in such a way that it fits in with both existing organizations’ ICTs infrastructure and operational environment, and whether the adoption of the model is supported by the top management of the organizations.

The analysis of the collected data shows that 60.0% (summation of 55.6% and 4.4%) of the respondents were satisfied with the organizations’ ICTs infrastructure to accommodate the model as shown in Fig. 6. It was also observed that 53.4% (summation of 46.7% and 6.7%) of the respondents were satisfied with the organizations’ operational environment to accommodate the model as shown in Fig. 7. Finally, the analysis results for this criterion show that 66.6% (summation of 53.3% and 13.3%) of the respondents were satisfied with the support provided by the top management of their organizations for the adoption of the model as shown in Fig. 8. These results carries the assurance that the model designs relatively can be applied in approximately 50 percent of the public organizations in the country.
4.1.6 Coverage and Completeness

Coverage and completeness is the extent to which something considers all parts or elements as originally planned. In designing a model, a developer should consider all aspects to make it complete and perfect. The aim of this evaluation criterion was to assess if the model was designed in such a way that it adequately addresses technical, non-technical, practice-related, and theory-related security issues.

The outcome of the analysis of the collected data shows that 80.0% (summation of 71.1% and 8.9%) of the respondents were satisfied with the technical security issues accommodated in the model as shown in Fig. 9. It was also observed that 75.6% (summation of 66.7% and 8.9%) of the respondents were satisfied with the non-technical security issues accommodated in the model as shown in Fig. 10. The results also show that 75.6% (summation of 68.9% and 6.7%) of the respondents percent of the respondents were satisfied with the practice-related security issues accommodated in the model. Finally, it was observed that 73.4 (summation of 66.7% and 6.7%) of the respondents were satisfied with the theory-related security issues accommodated in the model. Finally, it was observed that 73.4% (summation of 73.4% and 11.1%) of the respondents were satisfied with the theory-related security issues accommodated in the model as shown in Fig. 11. The results imply that the model designs in the proposed eGMM has considered all security related aspects to achieve a holistic secure model. Information security is a process; accordingly there is no equipment or tool that can replace the process. The reality is that no technical solution alone can make e-Government services more secure. This means the proposed eGMM is also open for improvement.

4.1.7 Compliance with Legal Aspects

Compliance means conforming to a rule, such as a specification, policy, standard or law. Security controls whether are technical or non-technical should be backed by policies, regulation, rules or laws. Unfortunately, information security policies and regulations if available in public organizations, they are not followed and most of them are outdated. The aim of this evaluation criterion was to assess if the model was designed in such a way that its implementation is supported by the country’s Acts, laws and regulations.

The outcome of the analysis of the collected data shows that 84.4% (summation of 73.3% and 11.1%) of the respondents were not satisfied with the
support provided by the country’s Acts, laws and regulations as shown in Fig. 12. The result confirms the fact that Tanzania lacks cyber laws that govern the protection of the online transactions. In general, security policies are applied to describe how organization plans to protect its ICTs assets. These plans should be supported by the country’s laws in case any of the organization member or citizen commits a cyber crime.

Figure 12: Respondents’ responses on the support of legal aspects to the implementation of the model

5. Recommendations

Based on the findings presented in this paper, we recommend the following activities to be done in order to improve both the model and security of e-Government services in the country:

1. Noting that, the presented results show that approximately 33 percent (one third) of the respondents expressed observation that the implementation of the proposed model in public organizations reduces the accessibility and usability of the e-Government services. For this reason therefore, we recommend that, the model be reviewed to provide a balance between achieving organizations’ information security and accessibility or usability of e-Government services.

2. Noting that, the results show that 40 percent of the respondents were not satisfied with organizations’ ICTs infrastructure to accommodate the model. We therefore recommend to target public organizations to improve their ICTs infrastructure in order to be able to use well the model for better security of e-Government services.

3. It was observed that approximately 47 percent of the respondents were not satisfied with organizations’ operational environments to accommodate the model. We recommend organizations to improve the condition of their operational environment. This improvement should also include the provision of an adequate budget to the organizations’ ICTs departments, and to support security awareness programs and training.

4. We recommend organizations to make more efforts in implementing both technical and non-technical security measures in order to protect e-Government services and gain citizens’ trust towards e-Government services.

5. Currently, Tanzania does not have specific legislations dealing with cyber security. There are no specific laws that govern and protect electronic transactions. For example illegal intrusion into a computer system cannot be prosecuted by the current legislations. Therefore, we recommend the government to hasten the enactment process of cyber laws to deal with data protection, cybercrime and electronic transactions, and e-Government services protection.

6. Conclusions

In this paper, evaluation of a holistic secured e-Government Maturity Model for protecting e-Government services in Tanzania is reported. A theoretical evaluation approach has been followed to test and validate the model. Specific evaluation criteria were selected and used to test and validate the model. The selected criteria include the following: simplicity, reliability, accessibility and usability, dynamics and flexibility, applicability, coverage and completeness, and compliance with legal aspects. Primary data were collected through questionnaires. The data were then processed and analyzed using the SPSS software. The overall evaluation result shows that the model designs meet all required specifications to successfully secure e-Government services, and the model is widely accepted by majority of the respondents at different organizational levels (strategic, tactical, and operational). Majority of the respondents expect that the model would enhance security by mitigating the current and future information security risks and threats posed to e-Government services. However, more efforts and time are needed to secure e-Government services properly. Further research work is recommended to test and validate the proposed model practically within the earlier studied organizations.
References


Authors Biographies

Mohamed D. Waziri is a doctoral student at the Nelson Mandela African Institution of Science and Technology. He holds a M.Sc. Computer Science from the University of Khartoum (2007). He also holds a B.Sc. in Computer Science from the International University of Africa (2003). Mr. Waziri is currently employed at the University of Dodoma (UDOM) as an Assistant Lecturer. Prior to joining UDOM he was working as an Assistant Lecturer at the International University of Africa, Khartoum, Sudan.

Eng. Dr. Zaipunna O. Yonah MIET, MIEE – holds a B.Sc degree (with Hons – 1985) in Electrical Engineering from University of Dar es Salaam – Tanzania; and M.Sc. (1986) and PhD (1994) Degrees in Computer-Based Instrumentation and Control Engineering from the University of Saskatchewan, Saskatoon – Canada. In Tanzania, he is a Registered Consulting Engineer in ICTs. Dr. Yonah has over 30 years of practice. His work spans the academia, industry and policy making fields. He is currently associated with The Nelson Mandela African Institution of Science and Technology – (school of Computation and Communication Science and Engineering), and the IEEE Inc. He is one of the pioneers driving the national broadband agenda in Tanzania. He believes that ICTs, as tools for development, promise so much: interactivity, permanent availability, global reach, reduced per unit transaction costs, creates increased productivity and value, jobs and wealth, multiple source of information and knowledge. Armed with such a belief, his current work aims at creating and delivering value through ICT enabled services in the shortest times possible. His research interest include: ICT4D, Cyber Security, ICT Policy and Regulation, Mobile and Web applications, high-capacity broadband networks, Intelligent Instrumentation and Control Engineering, and ICT enabled 21st Century Education delivery (ICT4E).
Load Sensitive Forwarding for Software Defined Networking – Openflow Based

Nor Masri Sahri¹ and Koji Okamura²

¹ Department of Advanced Information Technology,
Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan
normasri@kyudai.jp

² Research Institute for Information Technology, Kyushu University, Japan
oka@ec.kyushu-u.ac.jp

Abstract
Avoiding congestion is important role in recently proposed software defined network (SDN) since various new kind of overhead and delay introduced compared to traditional network. In this paper, we propose a load sensitive forwarding metric for the Openflow controller to make decision for path selection. Our metric assign weight based on the network traffic link load that put into account. This metric helps the forwarding protocol to load balance the network and improve the network capacity by avoiding the congested nodes to forward the traffic. Extensive performance results based on our simulation are presented to demonstrate the effectiveness of our proposed metric, with comparison to the SDN state of the art forwarding decision.

Keywords: Software Defined Networking, path selection, openflow

1. Introduction
In order to elaborate more on our proposed load aware forwarding metric, it is important to put the highlight on OpenFlow [1] which is a clean slate project introduced by Stanford University. OpenFlow was implemented as the first open standard interface for Software-Defined Network (SDN) architecture. SDN enable network administrators with a central programmable management interface, which is decoupled from the underlying network infrastructure for current layer 2 and layer 3 switches. In OpenFlow, the data path and the high level routing decision are made from two different devices, which is the OpenFlow-enabled switch and controller, respectively. The central controller provides the switches with the operational rules instructions, which is pushed by the controller to the switch as individual flow entries via a secured channel between them using OpenFlow protocol. The switches search the flow table corresponding entries and if there is any rules match, it will process the packets according to the pre-specified actions in the entries. The incoming packet in OpenFlow-enabled switches is matched against the flow table and the associated actions are taken into action: similar to the existing conventional switches. The main difference is that if the packet has not match any rule in the flow table, the packet would be drop or flooded in the network. OpenFlow-enabled switch include the packet header and encapsulate into the Openflow asynchronous message name Packet_In, and forwarded to the controller. The controller then use the flexibility of software to do analysis and further do the path selection decision based on shortest path algorithm. The controller then install a flow entry to the switch together with associated action via Packet_Out. The specification of the Openflow protocol message can be found in [2] for further clarification.

In other words, the Openflow controller has the helicopter view of a particular network and with this feature, the network management is expected to be much more “controllable” in single machine which is the controller itself. However, the SDN is also come with a various new type of overhead and cost. We identify a disadvantage of Openflow when compared to the native packet forwarding where all the unclassified incoming packet in a particular switch must be forwarded to the controller for further processing. Then the controller will decide the path for the packets and install the flow entry into the chosen Openflow switches flow table. The mechanism may introduce a significant delay and in order for the switches forward the incoming packet thru most optimal path, the controller also need to do the decision based on the current network state and condition which introduced another type of cost. The delay in a node with higher traffic load goes in and out through an interface, could be larger than a node with lower traffic load. If the controller include this...
heavy nodes as the flow entry to forward packet, it may actually cause end-to-end delay even longer. Furthermore, if one of the heavy nodes is congested, it may lead to packet drops and retransmission problem on certain nodes.

Many effort has been put into SDN research about load balancing and most notably that discussed load balancing in SDN is [3] where the author present algorithms that exploit the wildcard feature in Openflow that target to achieve equal fair distribution of the traffic and automatically adjust to changes without disrupting existing connections in data center network. In [4], the author proposed SDN load balancing that simplify the network placement in the network. When a new server is installed, the load balancing service will take appropriate decision to seamlessly distribute the network traffic among the available server, putting the network load and available computing capacity into consideration. They demonstrate that it can simplify the network management and provide some flexibility to network operators. Other recent proposals [5][6][7] also shows the load balancing work improvement in SDN.

In contrast to other works, we focus on designing a metric using Openflow protocol for controller to make path selection decision for forwarding a packet through most optimal available path. In this paper, we design a forwarding metric considering the available link load utilization that aim to minimize the network end-to-end delay. The delay through a node, which has larger link utilization could be larger than through the one, which has less traffic passing through their available ports.

The rest of the paper is organized as follows. Section 2 provides the details of our proposed load sensitive metric algorithm for SDN using Openflow protocol and in Section 3, we discuss some of our proposed metric implementation issues. In Section 4, we present the forwarding protocol design followed by the performance evaluation results in Section 5 and finally, Section 6 concludes this paper.

2. Load Sensitive Metric

We define network load over a path as the average rate of bytes that goes through the link and correctly received by the other node that attached to it. Therefore, the link load utilization is expressed as the current link usage over the maximum link capacity and is a number from zero to one. When the utilization is zero it means the link is not used and when it reaches one it means the link is fully saturated [8]. We use Load Sensitive for Software Defined Networking (LSSDN) terminology to denote the proposed load sensitive metric at each link.

To get the link load utilization, the formula are as follows

\[
\frac{\text{throughput}}{\text{datarate}} \times 100
\]

The throughput is measured by the sum of incoming and outgoing bytes. In Openflow, the needed information (byte count) is available from Openflow switch port counter [2]. To measure the link load utilization, the Openflow controller needs to monitor all switches port traffics, cache all of the number of incoming and outgoing bytes through their interface for further calculation.

Assuming that \( L \) links are available between two nodes. The bandwidth of \( i \)th link between node\(_a\) and node\(_b\) is \( B_{m-n}^i \) \((i=1,\ldots,L)\). If the total number of bytes going through an interface denote as \( T \), then we can calculate the traffic utilization as follows. We define the link load utilization of a node interface, \( U \) over \( link_i \) from node\(_a\) to node\(_b\) as the current total of throughput through the interface at certain time over the link transmission rate. The formula can be simplified as below.

\[
U_i = \left( \frac{\sum_{i=1}^{H} T_{m-n}}{B_{m-n}^i} \right) \times 100
\]

The path weight of the LSSDN metric is defined as (consider end to end path including \( H \) hops),

\[
\omega U_i = \sum_{i=1}^{H} U_i
\]

Note that the LSSDN metric given in (3) is under the assumption that all the packets can continuously go through all the path hop-by-hop without any node or link failure.

Let vector \([B_{m-n}^1, T_{m-n}^1, L]\) is characteristic of link between node\(_a\) and node\(_b\). \( B_{m-n}^i \) denote the bandwidth of \( i \)th channel between node\(_a\) and node\(_b\) \( L \) is the number of available links and \( T_{m-n}^i \) is the total throughput of \( i \)th link between node\(_a\) and node\(_b\). For a given network of \( G \) \((V, E)\), and the source node \( N_s \) and destination \( N_d \), the LSSDN algorithm include the following steps:

Step 1: Calculate the bandwidth capacity \( (B_{m-n}^i) \) for every link in network \( G \) \((V, E)\).
Step 2: Calculate the total throughput (in bytes) over a link $T_{m-n}$ for every link in network $G (V, E)$.

Step 3: Calculate the weight $\omega U_i$ for every link in $G (V, E)$ according to Eq. (3).

Step 4: Use Dijkstra Algorithm to find the smallest sum of weight in the paths of $G (V,E)$ from node $N_s$ to node $N_d$. The details of LSSDN are given in Algorithm 1.

**Algorithm 1: Smallest Link Utilization Path Selection Algorithm**

```
Input: $B_{m-n}, T_{m-n}, (i=1,\ldots,L)$;  
$V=\{v_1, v_2, \ldots, v_n\}$: The set of nodes;  
$N_s \in S$: source node;  
$N_d \in V$: destination node;  
for $j=1$ to $N$ do  
for $k=1$ to $N$ do  
for $i=1$ to $M$ do  
find $B_{m-n}^i$;  
find $T_{m-n}^i$;  
calculate $\omega U_i$;  
end for  
end for  
end for  
$S$: The current set of nodes (from $N_s$ to $N_d$) which has smallest load path  
$T(V_i)$: The current sum of link load of the links on the smallest weight path from $N_s$ to $N_d$;  
for $j=1$ to $N$ do  
$T(V_j) = \infty$;  
$T(N_d) = 0$;  
$S = \emptyset$;  
end for  
while $N_d \in S$ do  
\{  
u = v; // $N_i \in S$ and ($T(V_i)$ is smallest load in all nodes in $V \rightarrow S$)  
$S = S \cup \{u\}$;  
for all $v_k \in V \rightarrow S$  
if ($T(V_i) + \omega U_{m-n} \leq T(V_k)$)  
$T(V_k) = T(V_i) + T(V_k) + \omega U_{m-n}$;  
\}  
\}
```

2.1 Impact of traffic load utilization

Besides the update frequency, the number of transmitted and received bytes information affects the estimation of link utilization for the LSSDN metric. The number of throughput changes instantaneously. If we use the value directly for link utilization calculation, frequent rerouting might occurred. To avoid the problem, we maintain a weighted average link utilization in the controller, denoted as $\bar{U}$ and controller use this weighted average value as the backlog information instead of instantaneous sample value for the LSSDN computation. Specifically, the controller samples the instantaneous throughput according to a schedule, and let $U_n$ denote the $n^{th}$ sample. The average link load utilization, $\bar{U}$ by incorporating the instantaneous link utilization $U_n$ according to the exponential weighted moving average scheme [10], is

$$\bar{U} = (1 - \alpha)U_{old} + \alpha \cdot U_n \quad (4)$$

where $0 \leq \alpha \leq 1$ // $U_n = n^{th}$ sample

To show the need to include link utilization as a metric in SDN forwarding decision, we demonstrate the relation between link utilization and the input load using simple simulation. In the simulation, we tried to vary the aggregate input load traffic and measured the usage of the bandwidth. To simplify the network, the packet size was set to 1000 bytes and the link data rate was set to 1Mbps. The chart as illustrated in Fig. 1 is to emphasize the impact of the link load utilization on forwarding packet to their destination. From the figure, it clearly shows that the link load utilization is proportional to the traffic that sent through the link. It is observed that the bandwidth utilization is increase linearly with the input load and then it get saturated as the input load reach approximately 800 Kbps. When the link is saturated, the link load utilization is almost constant even though the input load increases and it is happened because the available link bandwidth is almost fully utilized. The higher the value of link load utilization, the lesser traffic can be send over the link.

Fig. 1 Link utilization as a function of input traffic

We also carried another simulation to study the capability of a link for to tolerate more traffic at different link load utilization. We simulate 5 pair of nodes exchanging data with another and we can derive the relation between link load utilization and
delay of packet. As illustrated in Fig. 2, we observed that the packet delay time increase dramatically when the link utilization start to climb at 90 percent of link utilized. This shows that the need to consider the link load is vitally important at high link utilization but its effect may also be ignored when the link load in under-utilized.

Fig. 2 Packet delay time as function of link utilization

3. IMPLEMENTATION DESIGN ISSUE

3.1 Update frequency

Our proposed LSSDN forwarding metric can be viewed as a load sensitive metric as it is heavily depend on the switch port information. Similar to other load sensitive metrics, the Openflow controller is require to perform recalculation by updating the traffic status to avoid usage of the congested link in the network. To balance the tradeoff between performance and the overhead, the route update frequency is a critical factor. More frequent updates of network state will introduce unnecessary overhead. On the other hand, large gap of update frequency will prevent the route from timely tracing the network status, and the network performance may dropped. We adopt the multipart message provided in Openflow feature which use to encode request or replies to or from the controller to switches. We simulate the feature to get port statistic for all of the switches to calculate our proposed link load metric in the controller. In our proposed algorithm, the message collect byte count information to measure the packets going in and going out through a particular port. In our simulations, we set the time for the Openflow switches to update the controller with the needed information automatically every 5 seconds. After the controller receive the new information it will perform recalculation based on our proposed metric to define the most optimal next forwarding path. In Openflow, the controller able to modify the existing flow entries action field that installed in the Openflow switches via flow table modification message that modify all flow that match. In our case, the controller will modify the existing flow entries in the switch with the newly most optimal next forwarding path.

4. FORWARDING PROTOCOL DESIGN

4.1 Route Discovery

We design a forwarding protocol for SDN IP network which aim to create congestion free flow entry by making use of information gathered from Openflow switches MAC layer. Now we describe the route discovery process in our proposed method. As illustrated in Fig. 3, any source node wishing to transmit data to a given destination will be process by the Openflow switch first. In Openflow, the switches contain exact matching tables for the forwarding databases. The incoming traffic will be check by the switches whether the packet has any matching in the flow table by performing a table lookup process. In our works, the matching field is based on the incoming source and destination IP address. When there is no match, the unmatched packet header will be extracted by the switch to be include in a special Openflow protocol message called Packet_In. This packet is use by the Openflow to transfer the control of any unmatched packet to the controller. We assume that all of the Openflow switches support the internal buffering to keep the unmatched packet in the switch buffer. Any unmatched packet will be buffered in the switch waiting to be forwarded to the next hop. Packet_In contain the buffer ID to represent the unmatched packet that stored and also some fraction of the packet header to be used by a controller when it is ready for the switch to forward the packet.
Any forwarding metrics require the real-time traffic information. When the controller receive the Packet_In message, it will send specially crafted type of message by Openflow to collect all of the switches port related statistic information via broadcast technique in order to decide the most optimal next hop to forward the packet to their destination. The message will request the number of transmitted and received bytes that go through all of the available ports of a node which is a vitally needed information for our proposed load sensitive metric. In Openflow, various kind of statistic information can be requested such as number of incoming and outgoing bytes or packets, the number of drop packets and also the time duration for how long the port has been alive in seconds.

4.2 Route Reply

After the switches receive the request message from the controller, it will include the needed current transmit and received bytes information from the ports in the port statistic reply message. The switches will send back the reply message together with the information back to the controller. Once the controller received the requested statistic, it will perform the path selection calculation to decide the most optimal next path for the switch to forward the packet using our proposed metric as presented in Section III. After the decision, it will install the flow entry into switches flow table via Packet-Out messages that contain the buffer ID referencing a packet that previously stored in the switch. The message also contain the action field that decide the next path for the switch to forward the buffered packet to their next destination. When the switch receive the Packet_Out message, it will install the flow entry in the current flow table and match the buffered packet with the identical buffer ID and continue to forward the traffic to the next path.

5. PERFORMANCE EVALUATION

The goal of our evaluation is to show the effectiveness of proposed metric to be adopted in SDN. We consider random topology which is based on Power Law model [9]. In this model, which is often use to represent the actual Internet, most of the nodes has lower number of links while a small number of nodes have a larger number of links. In our simulation, we set the number of nodes to 20 and the node degree is set to two which means that each nodes has at least four connected links on average.

Five nodes were randomly chosen to generate UDP traffic across the network, with the packet size of 512 Bytes respectively.

Fig. 3 and Fig. 4 present the results of our simulation that shows the performance of our proposed methodology with the comparison of the native Openflow forwarding mechanism. The total of network throughput and end-to-end delay versus the various flow rate transmission is chosen as the performance metric to be evaluated. The queue size is pre-defined to 20 packet in each switches and the simulation time is set to 600 seconds. From the two figures, it is explicitly demonstrated that our proposed metric result in much better performance for end to end delay than the native forwarding in Openflow under the random topology.

6. CONCLUSION AND FUTURE WORKS

In this paper, we evaluated the applicability of native OF data path selection for forwarding action in SDN environments. Our study shows that when the SDN network link is almost saturated, the packet delivery time is also increased hence the need to propose
forwarding path selection algorithm is proposed. We also demonstrate that by using our proposed metric, it can lead to path selection with minimum end-to-end delay and higher network throughput is also achieved. Since Openflow is a clean slate technology, various type of delay and network overhead is introduced. We plan to study those limitation and try to identify other types of possible metric for path selection problem in SDN network.

References


A Novel Vernier-based Time to Digital Converter for Low-power RFID Sensor Tags

Seyed Hossein Shahrokhi¹, Mehdi Hosseinzadeh²

¹ Department of Computer Engineering Science and Research branch, Islamic Azad University
Tehran, Iran
Sh.shahrokhi@srbiau.ac.ir

² Faculty Member of Computer Department, Science and Research Branch, Islamic Azad University
Tehran, Iran
Hosseinzadeh@srbiau.ac.ir

Abstract
Power consumption is a key factor in analogue and digital design for portable devices. Radio Frequency Identification (RFID) is widely used in industry, military and medical purposes. This technology operates with very low power consumption. Power consumption in passive tags is negligible and in fact, power consumption delimits the sensor tag’s life-time and range.

Many methods are proposed to reduce the RFID’s power consumption, but new researches are demanded due to technological advancements and application growth. In this research, a new scheme for time-to-digital converter was proposed which consumes a significant amount of energy in sensor tags. This method preserves accuracy and speed in measurement while reducing the area used by the circuit as well as power consumption by a significant amount.

Keywords: RFID, Low Power Design, Tag, Time-To-Digital Converter, Verinier, Sensor, TDC.

1. Introduction

Radio frequency identification (RFID) mostly indicates systems which transmit a subject’s ID number using a wireless technology. Barcode Technology was widely used before RFID for a touchless recognition, but now barcode readers cannot provide basic advantages of RFID. Although they are inexpensive, but they have some weaknesses, e.g., low capacity and unprogrammability.

Recently, combinations of RFID with different smart tags are pervasively used. This combination led to the creation of a new category of wireless sensor tags which emphasizes RFID’s importance and provides new applications [1]. Sensor tags traditional structure is based on generating a sensitive voltage or current which is sensitive to a physical magnitude and eventually converting it to a digital code by an analogue to digital converter. On the other hand, analogue current or voltage processing circuits face more challenges as physical dimensions shrink. For example, reducing the voltage supply level makes the converter more sensitive to noise or increasing the threshold voltage causes other challenges [2-3], [18].

In sensor tags, it occurs when the power supply is produced by voltage multiplier but the utilizable current dramatically reduces by distance. A fundamental scheme was proposed to change the algorithms from physical magnitude-based to time delay-based to overcome traditional restrictions in ADC-based sensors [4]. Afterwards, another scheme based on time-to-digital converter (TDC) was proposed [5]. Therefore, a time interval sensitive to a physical magnitude was created. This interval was converted to a digital code by a TDC. Different methods were proposed to generate a time interval sensitive to different physical quantities [6], [7]. TDCs have high accuracy and operate with low voltage, low power and low area consumption. Therefore, converting a physical magnitude to time using TDC is a good idea for RFID sensor tags. Basically, there are three methods to measure the time interval between two asynchronous signals: analogue, digital and interpolation.

The analogue method is usually based on the measurement of the changes of a capacitor voltage which is charged by an invariable current during a charging time interval. The analogue method is very accurate and unique but its weaknesses are low stability (high noise-sensitivity) and non-linearity [8]. Thus, it is only usable for short time intervals measurement.

The interpolation method in signal processing is based on digital filtering techniques. This method has acceptable accuracy and is more stable compared to the analogue method. Although it has some weaknesses like artificial
estimations and complicated calculations and it needs a significant time interval for appropriate functioning.

The digital method is based on the number of a reference oscillator's clock cycles. This method is linear on a long time interval and its power consumption is acceptable. Its structure is more appropriate to be implemented in integrated circuits. It is more reliable in terms of noise sensitivity compared to analogue methods, but still its accuracy is restricted to the oscillator's period. To increase the accuracy, the oscillator's frequency must be increased which increases power consumption, while the power consumption is a vital parameter in RFID tags [8].

In order to increase the accuracy and reduce power consumption, several researches have been carried out on TDCs in RFID [8], [9-15].

A simple digital TDC is formed by an oscillator and a counter. A Start signal initiates the oscillator and finally a Stop signal pauses it. The counter counts the number of oscillations and eventually time intervals between the Start and Stop can be found by the number of counter multiplied by oscillator period. In Fig. 1 counting the number of oscillator periods by a digital TDC is shown.

![Binary Counter](image)

This structure is simple but not accurate. As it is shown, the time inaccuracy is possible up to one oscillator period which is not acceptable in sensitive sensory tags.

The most important challenge of TDC design in RFID passive tags is to minimize the power consumption. A Sensor tag's general structure which uses a TDC to measure the time intervals is shown in Fig. 2 [1], [2].

![Sensor tag's general structure](image)

The analogue front-end, digital section and EEPROM are elements which are used in typical sensor tags. Sensing module is an extended block which is designed for RFID sensor tags. This block includes a magnitude to time converter section and TDC. The magnitude to time converter generates two signals which are Start and Stop. The duration between the start and stop rising edges ($T_d$) is a measurable magnitude. The Start and Stop signals are processed by TDC to generate a digital code. This code can be passed to the digital section and a sensed (measured) magnitude is sent to the reader. In the following section, the latest version of TDC will be introduced and compared with the other researches.

2. Vernier-based Time-to-digital Converter

Fig. 2 shows the general structure of Vernier-based TDC designed for RFID sensor tag. This method is also known as Nonius design [16, 2].

![Nonius TDC structure](image)

According to [17], measurement accuracy can be optimized using two oscillators and two counters.

In this section, new ideas of measuring the time interval with the aim of reducing the power consumption will be introduced. In fact, the proposed method is based on Vernier method. As shown above, in Vernier method, there are two oscillators, which the first one starts with the Start signal and the other starts with the Stop signal. Fig. 4 shows the time diagram for both oscillators. Signals 1 and 2 are the output signals for oscillators 1 and 2, respectively.

![Time graph in Vernier method](image)
Two oscillators and two counters are utilized in order to increase measurement accuracy. As it is shown in above, the time lapse between the Start and Stop signals is two to three times multiplied by Oscillator 1 period. But in order to recognize the non-integer value of counter N1, another oscillator with shorter period than oscillator 1 must be used. In this case, the non-integer value of counter N1 is reachable as well. When two oscillators are synchronized, it means both counters are showing integer values. The time interval between the Start and Stop signal can be estimated by:

\[ T_d = N_1T_1 - N_2T_2 \]  

Where T1 and T2 are signal 1 and 2 periods, respectively and T_d is exactly the time between Start and Stop which must be measured. Thus, in this case in order to measure this time interval, two oscillators and two counters are required.

### 3. The Proposed Design

#### 3.1 Signal Synchronization

Assume that we divide the time required for measurement (T_d) into two sections. As shown in Fig. 5, section \( \Delta T_1 \) which equals to the section dividable by T1 and section \( \delta 1 \) which is the remainder of T_d time. In fact, calculating this remainder section is of paramount importance, for if this section is ignored, there might be an error up to one period of signal 1 which is unacceptable for accurate sensing functions in RFID sensor tags.

\[ N_2 \times T_1 - N_1 \times T_1 = K \times T_1 \]

\( \Delta T_2 \) equals the time in which two oscillators are synchronized. That means \( \Delta T_2 \) is the time interval required for signal 2 to pass signal 1. In fact, this time interval is the required time to compensate for \( \delta 1 \) time interval.

Now, assume the state in which two signals are synchronized at the beginning and the end of \( \Delta T \) time interval, according to Fig. 6, the periods of signals 1 and 2 or T1 and T2 follow the following equations.

\[ T_1 = \Delta T/N_1 \]

\[ T_2 = \Delta T/N_2 \]

That means \( T_1/T_2 \) ratio equals \( N_2/N_1 \).

![Fig. 6 two signals synchronized at the beginning and the end of the time interval](image)

Assuming \( T_1 > T_2 \), it can be concluded that:

\[ N_2 = N_1 + K \quad , \quad (K \in Z>0) \]

\( N_2 - N_1 = K \)

This means signal 1 oscillated k times less than signal 2.

Multiplying signal 1 period by the equation's both sides will yield:

Because \( N_1 \times T_1 = N_2 \times T_2 \):

\[ N_2 \times T_1 - N_2 \times T_2 = K \times T_1 \rightarrow \]

\[ N_2 \times T_1 - N_2 \times T_2 \\ T_1 \]

\[ \rightarrow \]

\[ N_2 \times (T_1 - T_2) = K \]

The above function means if \( T_1/T_2 \) ratio is determined, in order to recognize the oscillation difference between the
two signals in this interval (k), only one cycle i.e., N2 is needed. Thus, in order to recognize K, both cycles counts (N1, N2) are not necessary.

3.2 Signals' Conditions in Time-to-digital Converter

In TDC, the conditions are different. As shown in Fig.7, the outputs of the two oscillators are not synchronized at the start of oscillator 2 which is concurrent to the Stop signal's edge. In fact, this phase difference makes the number of cycles in signal 1 non-integer. This is the value for measuring which the TDC is expanded and improved.

In \( \Delta T_2 \) interval, signal 2 oscillated exactly N2 times. According to two signals' periods ratio (T1/T2), the number of oscillations of signal 1 in this interval equals:

\[
\frac{T_2}{T_1} \times N_2
\]

Which is undoubtedly non-integer\(^1\) and smaller than N2. That means difference between N2 factor and 1, is the N2 factor in the non-integer portion before the Stop signal, that is:

\[
1 - \frac{T_2}{T_1}
\]

As it is shown in Fig.7, N2 is the number of oscillations which synchronizes the two signals. It means N2 is the compensator of non-integer portion of signal 1 in time interval \( \Delta T_2 \).

The non-integer fraction before the time interval (\( \Delta T_2 \)) is given by:

\[
\delta_{\text{n1}} = \frac{T_1 - T_2}{T_1} \times N_2
\]

Thus the number of oscillations of signal 1 in \( \Delta T_1 + \delta_1 \) time interval equals:

\[
N_0 + \delta N_1
\]

In which N0 is the integer number of signal 1 oscillations in the interval. Then, given this number multiplied by signal 1 period, the time in question can be calculated as:

\[
T_d = (N_0 + \delta_n) \times T_1
\]

The above equation means in order to measure the time between Start and Stop, we only need the number of oscillations of one signal. Thus, during the whole measuring time measurement of only one magnitude is necessary which leads to the omission of one counter in Vernier method. Thus, the system can be redesigned so that it contains only one counter.

3.3 The Circuit Design

The proposed TDC is shown schematically in Fig.8.

The circuit includes two oscillators, Main and Vernier. These two oscillators are of digital ring oscillator type. The NAND gates delay causes the oscillation in output loop. Each oscillator's frequency is a function of all gates delay and the number of the gates must be determined, considering the production technology. Fig.9 shows a NAND based ring oscillator structure.

Both oscillators outputs are passed to Oscillator Select module and one of them passes to the counter based on demand. In Fig. 10, an Oscillator Select module is shown. As the Stop signal is activated, the module's output switches from Main to Vernier.

---

\(^1\) In some T1/T2 ratios, signals 1 and 2 can never be synchronized.
A Phase Detector module is also used to determine the Main and Vernier synchronization moment. This moment is the end of measurement period. The Phase Detector module's structure is shown in Fig. 11.

The counter section counts the number of edges in its input signal.

3.4 Timing

As the Start signal is activated, the Main oscillator begins to act and its output is passed to the counter. When the Stop signal is activated, several tasks are done:

1. The counter’s content is stored in the next level’s latch which is the Digital Core.
2. Vernier oscillator begins functioning.
3. The counter is reset.
4. The oscillator select module switches and passes the Vernier oscillator’s output to the counter.

Figure 12 shows the time diagram for TDC circuit. As it is shown, the counter counts the number of the Main oscillator’s oscillations before the Stop signal and when the Stop signal is activated, it resets and begins counting the Vernier oscillator’s oscillations. The END signal is the Phase Detector’s output and indicates the end of the measuring process. As the END signal is activated, the counter stops and its content is stored in the Digital Core latch and both oscillators stop working.

Using this method, at the end of each measuring cycle, two numbers will be latched in the output one of which is the integer number of the Main oscillator’s oscillations before the Stop signal and the other one is the number of Vernier oscillator’s oscillations before the synchronization of its output signal with the Main oscillator. Using the 8-3 function, the time between the Start and Stop signals can be calculated.

The following diagram shows the output waveform for the circuit designed by VHDL code in which the designed circuit’s functionality is shown.

We designed the TDC circuit in a way that it only needs one counter. Omission of one counter reduces the area and power consumption in Nonius design [17] and we will investigate the new scheme’s circuit characteristics and compare them with the traditional designs.
4. Simulation

We utilized the Cadence software to simulate the circuit, and MOSFET 0.18um is the technology used in this simulation. Circuit's output waveform in 130ns Simulation is shown in Fig.14 to Fig.16. The simulation working frequency and input time interval are respectively assumed as 1 GHz and 10 ns.

4.1 Power Consumption

In table 1, the power consumption of the proposed method is compared to that of the traditional schemes which indicates about 14% reduction compared to Nonius design.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Conversion time</th>
<th>Power consumption (uW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSA TDC[17]</td>
<td>180 ns</td>
<td>900</td>
</tr>
<tr>
<td>Nonius TDC</td>
<td>130 ns</td>
<td>420</td>
</tr>
<tr>
<td>This work</td>
<td>130 ns</td>
<td>362</td>
</tr>
</tbody>
</table>

4.2 Layout and Area Consumption

The layout design for the proposed circuit is shown in Fig.17.

In table 2, the area consumption of the proposed method is compared to Nonius TDC. As it is manifest in the Table, the area consumption reduced about 25% in our method.
Table 2: Area consumption in the proposed method compared to Vernier method.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Area consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonius TDC</td>
<td>80 µm x 17 µm</td>
</tr>
<tr>
<td>This work</td>
<td>60 µm x 17 µm</td>
</tr>
</tbody>
</table>

5. Conclusion

RFID sensor tags are pervasively used. Thus, in this research, it was tried to investigate their performance improvement. The most important factor influencing performance is power consumption. The power consumption for typical RFID tags is about microwatts which show how important design improvement is in their circuits. In this paper, an improved design for time-to-digital converter was proposed. According to the simulation results, power consumption and area consumption overtake all traditional designs. In conclusion, this method can replace prevalent designs in RFID sensor tags.

References


Seyed Hossein Shahrokhi is PhD student in computer engineering at Islamic Azad University of Science and Research. He received MSc degree in computer engineering at Islamic Azad University of Science and Research Branch. His research interests are low-power design and Wireless Sensor Networks.

Mehdi Hosinzadeh is assistance professor at Islamic Azad University of Science and Research. He received his PhD in Islamic Azad University of Science and Research. His research interests are residue number system, computer networks, computer architecture and networks on chip.
Vulnerabilities and Improvements on HRAP\(^+\), a Hash-Based RFID Authentication Protocol

Seyed Mohammad Alavi\(^1\), Behzad Abdolmaleki\(^2\), and Karim Baghery\(^3\)

\(^1\) Imam Hossein Comprehensive University
Tehran, Iran
malavi@ihu.ac.ir

\(^2\) Information Systems and Security Lab (ISSL), Sharif University of Technology
Tehran, Iran
b.abdolmaleki.ir@ieee.org

\(^3\) Information Systems and Security Lab (ISSL), Sharif University of Technology
Tehran, Iran
k.baghery.1988@ieee.org

Abstract
In the last decade, Radio Frequency Identification (RFID) systems are employed in many authentications and identifications applications. In RFID systems, in order to provide secure authentication between RFID users, different authentication protocols proposed. In 2011, Cho et al. proposed a hash-based mutual RFID authentication protocol (HRAP). They claimed that HRAP protocol provides secure communication between RFID users and also it can provide users privacy. In that year, Habibi et al. investigated the security and privacy of HRAP protocol and showed that HRAP protocol has some weaknesses. Then, Habibi et al. proposed an improved version of HRAP protocol (HRAP\(^+\)) that eliminates all weaknesses of HRAP protocol. In this study, we cryptanalyze the HRAP\(^+\) protocol and we show that there are some flaws in HRAP\(^+\) protocol still. It is shown that, an attacker can perform tag impersonation, server impersonation, and replay attacks with success probability greater than \(\frac{1}{4}\). Then, in order to omit all mentioned weaknesses, we propose an improved version of HRAP\(^+\) protocol. Security analysis shows that the improved protocol can improve the performance of HRAP\(^+\) protocol. In addition, we compare the security of the proposed protocol with some hash-based protocols that proposed recently.

Keywords: RFID authentication protocols, HRAP\(^+\) protocol, Security, Impersonation Attack.

1. Introduction
RFID systems are increasingly becoming part of our daily life. In many of our daily routines, without realizing it, we use RFID technology that use radio waves for automatic identification applications [1]. RFID technology also is used in different objects for different applications. Mainly, RFID systems consist of three main parts including Tag, Reader and Back-end-server or Database (Shown in Fig. 1). The data included in RFID tags that often are identification numbers, can be collected by the wireless reader. Also the reader can perform some logic processors and change the content of RFID tags. The third part of RFID systems that contains all secret information of tags, is back-end server or database. The reader located between the tags and the back-end server and exchanges data between them. In each run of RFID system, the database performs some certification and authentication processes and provides access to the data [2]. In some applications, communication channels between the readers and the database is insecure [3]. But in some cases, communication channels between the readers and the database is secure [4].

Due to nature of wireless communication between tags and readers, these channels can be eavesdropped by an adversary. As a result, although these systems provide many useful services, they can dangerous for security and the privacy of end-users. In the last few years, in order to protect RFID users against different security and privacy attacks

![Fig. 1. A System model of RFID systems](image-url)
and provide secure communication between them, different RFID authentication protocols have been proposed [5-10]. Although, all proposed protocols have been presented to provide security and privacy of end-users, in some cases it is shown that the proposed protocols have some weaknesses and suffer from various attacks. So in order to increase the security and privacy of the proposed protocols, lots of literature focused in cryptanalyze of RFID authentication protocols [5-14].

In 2011, in order to provide secure communication for RFID users, Cho et al. proposed a hash-based mutual RFID authentication protocol [5] which referred as HRAP protocol in this paper. In HRAP protocol, communication channel between the reader and the database is secure. Cho et al. analyzed the security and the privacy of HRAP protocol and claimed that their protocol can provide security and privacy of RFID users. In that year, Habibi et al. [6] cryptanalyzed HRAP protocol and showed that still the security and the privacy of HRAP protocol has some problems and is not secure against desynchronization attack, traceability and backward traceability attacks. Then, Habibi et al. applied some changes in the structure of tag message ($M_1$) and proposed an improved version of HRAP protocol (HRAP+). Habibi et al. present some security and privacy analysis for HRAP protocols and claimed that HRAP+ protocol eliminates all weaknesses of HRAP protocol and is resistance against various attacks.

In this study, we cryptanalyze the HRAP+ protocol and we show that although Habibi et al. tried to omit all weaknesses of HRAP protocol, still HRAP+ protocol has some security problems and is vulnerable against tag impersonation, server/reader impersonation and replay attacks. In the HRAP+ protocol, the structure of RID has a problem that makes it vulnerable against the mentioned attacks. In this paper, it is shown that how an attacker can use this weakness and impersonate the tag, the back-end server or the reader. Mentioned attacks are based on an assumption that is reasonable in many cases. Due to this assumption, the success probability of mentioned attacks is greater than $\frac{1}{4}$ that are given in the section 3 with more details.

Furthermore, in order to increase the performance of HRAP+ protocol and provide security and privacy of RFID users, we propose an improved version of HRAP+ protocol. We analyze the security of the proposed protocol and we show that with our modifications all weaknesses of HRAP+ protocol removed. Also we compare the security of proposed protocol with some hash-based protocols that proposed recently. Our comparisons, show that the proposed protocol has sufficient security and privacy and is resistance against all attacks.

The rest of paper is organized as follows: HRAP+ protocol is introduced in section 2. In section 3, some attacks on HRAP+ protocol presented. In section 4, we apply some changes in HRAP+ protocol and propose an improved version of it. The security of proposed protocol is analyzed in section 5, and it is shown that all weaknesses of HRAP+ protocol are omitted. Also in this section the security analysis of proposed protocol are compared with some similar protocols that are hash-based and proposed in recent years. Finally, we conclude this paper in section 6.

<table>
<thead>
<tr>
<th>Server / Reader $(ID_{old}, S_{old}, ID_{new}, S_{new})$</th>
<th>Tag $(ID_i, S_j)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generates $\beta$ and obtains $R^i_k$ and $RID_i$</td>
<td>Generates $R^i_t$ Randomly: $R^i_t = (R^i_t - R^i_t mod S_j + 1)_{[0-47]}$</td>
</tr>
<tr>
<td>Calculates $\theta = h(\beta_i</td>
<td></td>
</tr>
<tr>
<td>If $ID = ID_{new}$</td>
<td>$\parallel (R^i_t + S_j - R^i_t mod S_j)_{[48-95]}$</td>
</tr>
<tr>
<td>$S_{old} \leftarrow S_{new} \leftarrow h(S_{new} \parallel RID_i)$</td>
<td>$\alpha_i = h(ID_{old} \parallel R^i_t \parallel R^i_t \parallel RID_i)$</td>
</tr>
<tr>
<td>$ID_{old} \leftarrow ID_{new} \leftarrow h(ID_{new} \parallel S_j)$</td>
<td>$\beta_i = ID_{old} + S_j \parallel S_j_{[0-47]}$</td>
</tr>
<tr>
<td>If $ID = ID_{old}$</td>
<td>$\theta \leftarrow h(\beta_i \parallel RID_i)$</td>
</tr>
</tbody>
</table>
| $S_{new} \leftarrow h(S_{new} \parallel RID_i)$ | If ($\theta == h(\beta_i \parallel RID_i)$) server is legitimate and the tag updates:
| $ID_{new} \leftarrow h(ID_{new} \parallel S_j)$ | $S_{j+1} \leftarrow h(S_j \parallel RID_i)$ |

Fig. 2. The HRAP+ protocol [6].
2. The HRAP\(^+\) Protocol

In [6], Habibi et al. proposed an improved version (HRAP\(^+\)) of HRAP protocol that proposed by Cho et al. in [5]. HRAP\(^+\) protocol is similar to HRAP protocol and consists of three phases. The structure of HRAP\(^+\) protocol is illustrated in Fig. 2. The notation that are used in HRAP\(^+\) protocol are provided in Table 1.

Table 1. The Notations of HRAP\(^+\) Protocol

<table>
<thead>
<tr>
<th>Notations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S_j)</td>
<td>The communications key shared by server and tag</td>
</tr>
<tr>
<td>ID(_k)</td>
<td>The group identification code of the (k) th tag</td>
</tr>
<tr>
<td>(R_t)</td>
<td>A random number</td>
</tr>
<tr>
<td>(h(.))</td>
<td>Hash function</td>
</tr>
<tr>
<td>(\parallel)</td>
<td>Concatenation operation</td>
</tr>
<tr>
<td>(A \oplus B)</td>
<td>Message A is XORed with message B</td>
</tr>
<tr>
<td>(A \not\equiv B)</td>
<td>Compare whether A is equal to B or not</td>
</tr>
</tbody>
</table>

3. Security Analysis of HRAP\(^+\) Protocol

In this section, the security of HRAP\(^+\) protocol is analyzed. It is shown that security of HRAP\(^+\) protocol has some weaknesses and does not resist against replay impersonation, reader impersonation and replay attacks. All attacks performed with two assumptions and the success probability of attacks are greater than \(\frac{1}{4}\) that in the rest of paper will be explained with more details.

According to the authentication phase of HRAP\(^+\) protocol, the value of \(RID\) is defined as follows,

\[
RID = (R_t - R_t \mod S_j + 1) \parallel (R_t + S_j - R_t \mod S_j)_{[0:47]} \parallel (S_j)_{[48:95]}
\]

(1)

now if \(R_t < S_j\), the value of \(RID_t\) in (1) can be rewritten as follows,

\[
RID = (R_t - R_t + 1) \parallel (R_t + S_j - R_t)_{[48:95]}
\]

(2)

\[
= (1)_{[0:47]} \parallel (S_j)_{[48:95]}
\]

Note that, when \(R_t < S_j\), \(R_t \mod S_j = R_t\).

3.1 Tag Impersonation Attack

This attack can be performed in two phases as follows.

**Learning phase:** In round \(i\) th of protocol, the attacker eavesdrops exchanged data between the tag and the server and obtains \(R_t^i\), \(R_t^i \oplus \beta\) and \(\alpha\).

**Attack phase:** In round \((i + 1)\) th of protocol, when the server send a request message \(R_t^{i+1}\), the attacker impersonate the tag and responses with \(\alpha\) and \(R_t^i \oplus \beta \oplus R_t^i \oplus R_t^{i+1}\) to the server. Then, the server performs following operations,

- For each tuple of \((ID_k,S_j)\), the server generates \(\beta\) and obtains \(R_t^{i+1} = R_t^i \oplus R_t^f \oplus R_t^{i+1}\) and \(RID_{i+1}\).
- Then the server uses the values \(ID_k^i, R_t^{i+1}\) and \(RID_{i+1}\) and checks that \(\alpha \gg h(ID_k^i \oplus R_t^{i+1} \oplus R_t^i \oplus RID_{i+1})\). According to mentioned assumptions \(R_t^i < S_j\) and \(R_t^{i+1} < S_j\), we can write \(RID_{i+1} = (1)[0:47] \parallel (S_j)_{[48:95]} = RID_i\).
- Since \(RID_{i+1} = RID_i\), the server authenticates the attacker as a legitimate tag.

**Proof:**

\[
h(ID_k^i \oplus R_t^{i+1} \oplus R_t^i \oplus RID_{i+1}) = h(ID_k^i \oplus R_t^i \oplus R_t^i \oplus RID_{i+1})
\]

\[
= h(ID_k^i \oplus R_t^i \oplus R_t^i \oplus RID_i) = \alpha
\]

(3)

**Lemma 1:** For two numbers \(\lambda\) and \(\mu\) that are random numbers from set \(\chi = \{0,1,\ldots,2^n - 1\}\) and \(\mu > 1\), the probability of inequality \(\lambda < \mu\) is greater than \(\frac{1}{2}\).

**Proof:** Provided in appendix.

**Lemma 2:** For three numbers \(\lambda\), \(\mu\) and \(\kappa\) that are random numbers from set \(\chi = \{0,1,\ldots,2^n - 1\}\) and \(\mu > 1\), the probability of inequality \(\lambda \oplus \kappa < \mu\) is greater than \(\frac{1}{2}\).

**Proof:** Provided in appendix.

According to the **Lemma 1** and **Lemma 2**, the inequalities \(R_t^i < S_j\) and \(R_t^{i+1} < S_j\) hold with probability greater than \(\frac{1}{2}\). Therefore, this attack will be successful with probability greater than \(\frac{1}{4}\).

3.2 Server Impersonation and Reply Attacks

In this section, we aim to show that in HRAP\(^+\) protocol, an attacker can perform replay attack and impersonate the server. This attack can be summarized as follows,

- Firstly, the attacker eavesdrops first session of protocol and obtains \(h(\beta \parallel RID_i)\). Also in this session, the attacker blocks third phase of protocol (transmit message from the server to the tag). As a result the tag does not update its secret values.
- Now, the attacker acts as a legitimate server and sends a random number \(R_{ATC}\) to the target tag.
In this attack also, according to the \( \mathcal{R} \) protocol, we propose an improved version of HRAP hash function. The improved protocol can be summarized as follows:

\begin{itemize}
  \item In response, the tag generates a random number \( R^t+1 \), and calculates \( R^t+1 \oplus \beta_{i+1} \) and \( \alpha_{i+1} = h(ID^t \oplus R_{\text{old}} \oplus R^t+1 \oplus RID_{i+1}) \), then sends them to the attacker.
  \item Then, the attacker sends eavesdropped message \( h(\beta_{i} \parallel RID_{i}) \) to the target tag.
  \item Since the tag dose not its secret values, \( \beta_{i+1} = \beta_{i} = ID_{k[48:95]} \parallel S_{j[0:47]} \). Using assumptions \( R^t_i < S_j \) and \( R^t+1_i < S_j \), it can be result that \( RID_{i+1} = (1)_{[0:47]} \parallel (S_j)_{[48:95]} = RID_{i} \). As a result \( h(\beta_{i+1} \parallel RID_{i+1}) = h(\beta_{i} \parallel RID_{i}) \) the tag authenticate the attacker as a legitimate server.
\end{itemize}

In this attack also, according to the Lemma 1 and Lemma 2, the inequalities \( R^t_i < S_j \) and \( R^t+1_i < S_j \) hold with probability greater than \( \frac{1}{2} \), as a result the attacker will impersonate the server with probability greater than \( \frac{1}{4} \).

4. Improved Version of HRAP\(^{+}\) Protocol

In the last section we showed that due to structure of \( RID_{i} = R_{t} - R_{t} \mod S_{j} + 1 \mid_{[0:47]} \parallel \big(R_{t} + S_{j} - R_{t} \mod S_{j}\big)_{[48:95]} \), an attacker could perform tag impersonation, reply attack, and server impersonation attack on HRAP\(^{+}\) protocol. In this section, in order to omit mentioned weaknesses of HRAP\(^{+}\) protocol, we propose an improved version of HRAP\(^{+}\) protocol (Shown in Fig. 3). In the improved protocol, we changed the structure of \( RID \), indeed we protect \( RID \) via a one-way hash function. The improved protocol can be summarized in two phases as follows.

\subsection{4.1 Initial Phase}

In this phase, some secret values such as \( ID_{k} \) and \( S_{j} \) are stored in the specific tag. Also, a one-way hash function is saved in all tags. In the server, for each specific tag, the values \( ID_{\text{old}}, ID_{\text{new}}, S_{\text{old}} \) and \( S_{\text{new}} \) are stored. Like as all tags, the server uses a one-way hash function in authentication procedures.

\subsection{4.2 Authentication Phase}

The authentication of proposed protocol is similar to HRAP\(^{+}\) protocol and consists of three phases. This phase can be expressed as follows.

1. Like as HRAP\(^{+}\) protocol, the server generate a random number \( R^t_r \) and sends it to the target tag.
2. Firstly, the tag generates random number \( R^t_i \). Then the tag uses \( R^t_i \) and calculates messages \( RID_{i} \). If they were the same, the server computes \( \alpha_{t} = h(ID_{k} \oplus R_{t} \oplus R^t_i \oplus RID_{i}) \).
3. In order to authenticate the tag, the server performs following operations.
   - For each tuple of \( (ID_{k}, S_{j}) \), the server generates \( \beta \) and obtains \( R^t_i \) and \( RID_{i} \).
   - Then the server uses the values \( ID_{k}, R^t_i, R^t_i \) and \( RID_{i} \) and checks that \( \alpha_{t} = h(ID_{k} \oplus R_{t} \oplus R^t_i \oplus RID_{i}) \). If they were the same, the server computes \( \theta = h(\beta_{i} \parallel RID_{i}) \) and sends it to the target tag.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
\textbf{Server / Reader} \((ID_{\text{old}}, S_{\text{old}}, ID_{\text{new}}, S_{\text{new}})\) & \textbf{Tag} \((ID_{k}, S_{j})\) \\
\hline
\multicolumn{2}{|c|}{\textbf{For each tuple of}} \((ID_{\text{old}}, S_{\text{old}})\) \textbf{and} \((ID_{\text{new}}, S_{\text{new}})\) \textbf{generates} \( \beta \) \textbf{and obtains} \( R^t_i \) \textbf{and} \( RID_{i} \) \textbf{Verify} \( \alpha_{t} = h(ID_{k} \oplus R_{t} \oplus R^t_i \oplus RID_{i}) \) \textbf{Calculates} \( \theta = h(\beta_{i} \parallel RID_{i}) \) \textbf{and sends it to the tag and updates its secret values as follows:} \\
\hline
\text{If} \( ID = ID_{\text{old}} \) & \( S_{\text{old}} \leftarrow S_{\text{new}} \leftarrow h(S_{\text{new}} \parallel RID_{i}) \) \\
\hline
\text{If} \( ID = ID_{\text{new}} \) & \( ID_{\text{old}} \leftarrow ID_{\text{new}} \leftarrow h(ID_{\text{new}} \parallel S_{j}) \) \\
\hline
\text{Generates} \( R^t_i \) \textbf{Randomly} & \( \alpha_{t} = h(ID_{k} \oplus R_{t} \oplus R^t_i \oplus RID_{i}) \) \\
\hline
\textbf{Calculate} \( \beta_{i} = h(ID_{k}[48:95] \parallel S_{j}[0:47]) \) & \( \theta \) \textbf{and} \( h(\beta_{i} \parallel RID_{i}) \) \textbf{server is legitimate and the tag updates:} \\
\hline
\text{If} \( \theta = h(\beta_{i} \parallel RID_{i}) \) & \( S_{j+1} \leftarrow h(S_{j} \parallel RID_{i}) \) \\
\hline
\text{ID}_{i+1} \leftarrow h(ID_{i} \parallel S_{j}) \) & \( ID_{i+1} \leftarrow h(ID_{i} \parallel S_{j}) \) \\
\hline
\end{tabular}
\caption{Improved version of HRAP\(^{+}\) protocol.}
\end{table}

54

Copyright (c) 2014 Advances in Computer Science: an International Journal. All Rights Reserved.
After that, the server updates its secret values similar to HRAP*, otherwise aborts the protocol.

- The tag calculates \( h(\beta_i \parallel RID_i) \) and compares with received message from the server. If they were the same, the tag authenticate the server and updates its secret values similar to HRAP* protocol.

5. Security Analysis of Proposed Protocol

In the last session we proposed an improved version of HRAP* protocol that removes mentioned weaknesses. In this subsection, we aim to analyze the security and the privacy of proposed protocol against various attacks.

5.1 Tag and Server Impersonations Attacks

In section 3, we showed that the main weakness of HRAP* protocol is the structure of \( RID = (R_t - R_i) \mod S_j + 1 \) that makes HRAP* vulnerable against impersonation attacks. In proposed protocol, we changed the structure of \( RID \) completely as follows,

\[
RID_{\text{new}} = h(R_t \oplus S_j)
\]

where \( h(\cdot) \) is a one-way hash function. As it can be seen, with this changes, since the values of \( R_t \) and \( S_j \) change in each run of protocol, and the attacker dose not access to them directly, so the attacker cannot perform impersonation attacks.

5.2 Replay Attack

In this attack, the attacker tries to perform impersonation attacks to access exchanged messages, modify, and even delete them. In the proposed protocol, we applied some changes in the structure of exchanged data between the tag and the reader, indeed we changed \( \beta = ID_k \parallel S_j \) to \( = h(ID_k \parallel S_j) \). Also, we changed the structure of secret value \( RID \) that provided in (4). It can be seen that with these changes, the attacker cannot perform replay attack. Note that, with new values of \( \beta \) and \( RID \), if somehow the attacker obtains the random number \( R_t \), he/she cannot extract secret keys \( S_j \) and \( ID_k \).

Furthermore, the structure of the proposed protocol is similar to HRAP* protocol, as a result the proposed protocol is secure against other attacks like as HRAP* protocol. More analysis about other attacks provided in [6].

In order to more evaluation of the security and the privacy of the proposed protocol, in Table 2, the security and the privacy of proposed protocol compared with some hash-based protocols that proposed in the last few years. It can be seen, that with applied new changes in the proposed protocol, all weakness of HRAP* protocol have been omitted.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag Impers.</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Replay Attack</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Reader Impers.</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>DoS Attack</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Traceability</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

✓: Secure  ×: Insecure

6. Conclusions

In this paper, the security of HRAP* that is an improved version of HRAP protocol, is analyzed. We showed that although the designer of HRAP* tried to remove all weaknesses of HRAP protocol, still HRAP* protocol has some security problems and is not resist against tag impersonation, server impersonation and replay attacks. Mentioned attacks were based on an assumption that in many cases is reasonable. Due to this assumption, the success probability of mentioned attacks was greater than \( \frac{1}{4} \). Furthermore, we presented an improved version of HRAP* protocol that removed weaknesses of HRAP* protocol. In order to more evaluation we analyzed the security of proposed protocol and also we compared the security analysis of the proposed protocol with some hash-based protocol that are in the same family and proposed recently.

Appendix

Proof of Lemma 1:

\[
A := \{(\lambda, \mu)|0 \leq \lambda \leq 2^n, 2 \leq \mu \leq 2^n, \text{and } \lambda < \mu\}
\]

\[
= \bigcup_{\mu=2}^{2^n-1} \{\lambda, \mu)|\lambda = 0,1, \ldots, \mu\},
\]

and

\[
S := \{(\lambda, \mu)|0 \leq \lambda \leq 2^n, 2 \leq \mu \leq 2^n\}
\]

as a result,

\[
Pr[\lambda, \mu \in A] = \frac{|A|}{|S|} = \frac{2^n(2^n+1) - 3}{2^n(2^n-2)}
\]
Proof of Lemma 2:
\[
A(\lambda, \mu) := \{(\lambda, \mu, \kappa) | 0 \leq \psi \leq \mu \}
\]
Thus, for each \((\kappa, \lambda, \mu) \in A(\lambda, \mu)\) there exist an \(\psi\) such that
\[
\lambda \oplus \kappa = \lambda \oplus \lambda \oplus \psi = \psi \leq \mu
\]
Now, let
\[
A := \bigcup_{0 \leq \lambda \leq 2^n, 2 \leq \mu \leq 2^n} A(\lambda, \mu)
\]
Hence, for total numbers \((\lambda, \mu, \kappa)\) such that \(\lambda \oplus \kappa \leq \mu\) and \(\mu > 1\) we have
\[
|A| = \sum_{\mu=2}^{2^n-1} \sum_{\lambda=0}^{2^{n-1}} |A(\lambda, \mu)| = \sum_{\mu=2}^{2^n-1} \sum_{\lambda=0}^{2^{n-1}} (\mu + 1)
\]
\[
= 2^n \sum_{\mu=2}^{2^n-1} (\mu + 1)
\]
\[
= 2^n \left( \frac{2^n(2^n + 1)}{2} - 3 \right)
\]
On the other hand,
\[
S := \{(\kappa, \lambda, \mu) | 0 \leq \lambda, \mu \leq 2^n - 1, 2 \leq \kappa \leq 2^n - 1\}
\]
\[
|S| = 2^n 2^n (2^n - 2)
\]
Now, the probability that for random \(\kappa, \lambda\) and \(\mu > 1\), \(\lambda \oplus \mu < \kappa\) is equal with
\[
Pr\left[(\kappa, \lambda, \mu) \in A\right] = \frac{|A|}{|S|} = \frac{1}{2} + \frac{3}{2(2^n - 2)} - \frac{3}{2^n (2^n - 2)} > \frac{1}{2}
\]

References


Analysis of Inter cluster movement based on geometric probability and regression

Rajee.A.M¹, Sagayaraj Francis.F²
¹ Department of CSE, Pondicherry Engineering College
Puducherry 605014, India
rajee.am@gmail.com

² Department of CSE, Pondicherry Engineering College
Puducherry 605014, India
fsfrancis@pec.edu

Abstract
This paper proposes an approach to study the behavior of clustering system, when unclustered data comes to the existing clustered setup. With the introduction of a new data point to the system, there is a possible occurrence of migration of points between clusters, resulting in re-clustering of the setup. An attempt was made to assess the inter cluster movement scenario, with geometric probabilistic approach and regression model, studying the possibility of migration initiated by the new data point which may be located at varying distance from the cluster center and simultaneously increasing the size of the clusters and dimension of the clustering data. A comparative study on the results obtained by both these models was framed and the results reveal lower relative error, which increases with increasing size and the dimension of the clustering system.

Keywords: Data clustering, Geometric probability, Inter cluster movement, Probability of Migration, Regression Analysis.

1. Introduction
Clustering is the process of organizing objects into groups, whose members are similar in some way. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. Cluster member are those data points belonging to the cluster.

Consider a clustered system built with well separated K clusters C₁, C₂, -----, Cₖ. The system is static when no additional data enters the clustered setup. Suppose, when a new entrée comes, the centroids of the cluster gets changed and other data members gets reassigned to the nearest cluster, thereby signifying in the possible movement of data points between clusters. This is known as inter cluster movement or inter cluster migration.

Technically, with the induction of a new input at d Euclidean distance to any of the closest cluster centers c₁, c₂, ---, cₖ, there may be possible movement of data from one cluster to another, which is obviously triggered by the new entrée. An approach to predict this occurrence of inter cluster migration will result in the reduction of number of times re-clustering is done on the dataset. This paper uses the concept of geometric probability along with prediction analysis using regression to predict the aforementioned cluster dynamics. This approach gives a special zest to the clustering system analysis, with a different diversified enhancement of the system.

The terms inter cluster movement and inter cluster migration stated throughout this paper proposes the movement of data points between clusters. The term distance mentioned throughout this paper represents the Euclidean distance of the new point to the closest cluster center cₖ. A cluster Cᵢ is closest (or synonymously nearest) to cluster Cⱼ if the Euclidean distance between their centers is smallest among each of the cluster pairs in the clustering system. We therefore refer Cᵢ is closest to Cⱼ and vice versa.

2. An overview of Geometric probability and regression analysis

2.1 Geometric Probability
A probability is a number from 0 to 1 that represents the chance that an event will occur. Assuming that all outcomes are equally likely, an event with a probability of 0 cannot occur. An event with a probability of 1 is certain to occur, and an event with a probability of 0.5 is just as likely to occur as not.

Geometric probability is the likelihood of an event occurring based on geometric relationships such as area, surface area, or volume.
Geometric probability is better demonstrated by its practice in game application. A common game is darts.

2.2 Regression Analysis

Regression analysis uses data to identify relationships among variables and use these relationships to make predictions. The variable that is to be predicted \((Y_c)\) is called the dependent (or response) variable. The variable \(X\) is called the independent (or predictor, or explanatory) variable. The simple regression model is based on the equation for a straight line: \(Y_c = A + BX\)

Where:

\(Y_c\) = The calculated or estimated value for the dependent (response) variable
\(A\) = The Y intercept, the theoretical value of \(Y\) when \(X = 0\)
\(X\) = The independent (explanatory) variable
\(B\) = The slope of the line, the change in \(Y\) divided by the change in \(X\), the value by which \(Y\) changes when \(X\) changes by one.

For a given data set, \(A\) and \(B\) are constants. They do not change as the value of the independent variable changes.

3. Related works

Regression analysis is used in clustering particularly in K Means in finding the appropriate number of clusters and testing the efficacy of K Means and other clustering algorithms. Various data sets including time series, heterogeneous data sets were used for clustering and subsequently regression parameters were involved to validate the correctness of the application. Correlation of regression analysis with clustering techniques was adopted by many researchers and few are as presented here.

Hongxing He, Jie Chen, Huidong Jin and Shu-Heng used K Means Clustering algorithm to partition stock price time series data[1]. After clustering, linear regression is used to analyze the trend within each cluster. Hammouda used simple regression technique and tested the accuracy and performance of four different clustering algorithms on various regression parameters including root mean squared error, regression line slope values[2]. Qian and Wu proposed a new regression based algorithm to determine the number of clusters and other underlying regression parameters[3]. Geeta, Moin and Arvinder considered heterogeneous software engineering data sets and classified into different clusters[4]. A combination of clustering and regression techniques can be used to reduce the potential problem in effectiveness of predictive efficiency due to heterogeneity of the data.

Emre, Ijker and Murat executed simulation model of a system and using K Means algorithm which created a regression model which are known as meta models for each cluster[5]. This approach increases the accuracy of clustering meta model, thereby decreasing sum of squared value. Qin and G.Self proposed a new clustering method, the clustering of regression models (CORM) method employs regression to model gene expression and clusters genes based on their relationship between expression levels and sample covariates[6].

4. Implementation of geometric probability in predicting inter cluster movement

K clusters \(C_1, C_2, ..., C_K\) with corresponding centers \(c_1, c_2, ..., c_K\) were built and presented as circles for this study. A new data point was introduced with distance \(d\) around \(c_k\). The new point may either be located as shown in figures 1a, 1b, 1c, 1d or at an arbitrary distance around \(c_K\). The chance of any new point within the boundary of \(C_k\) effecting inter cluster migration is apparently nil. Let \(P(d)\) be the probability of migration when a new data point is placed at a distance \(d\). Let \(P = \{P(d), P(d+\delta), P(d+2\delta), ..., P(d+n\delta)\}\) signifies the probability of migration with respect to \(\delta\) increase in distance \(d\).
Figures 1a, 1b and 1c signify the data point located at increasing distance $d$ from its closest cluster center. The solid circle represents an outermost boundary of the cluster. The dotted circle signifies the possible location of the data point at increasing distance from its closest cluster center.

A data point introduced at a distance $d'$ as shown in Fig 1d, causes movement of points between the clusters. Hence at $d'$, the probability of inter cluster migration is 1. Below $d'$, the probability ranges from $0 \leq p < 1$. A zero probability is achieved when the new data point comes very close to $c_1$.

The geometric probability of occurrence of inter cluster movement is calculated as in (2).

Applying (2), with increasing distance $d$, we get corresponding probability values as in Table 2.

Table 1: Probability of inter cluster movement at varying distance of the data point

<table>
<thead>
<tr>
<th>Distance $d$ from cluster center $c_1$ $(X_i)$</th>
<th>Probability of Migration $(in % )$ $(Y_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.082195</td>
</tr>
<tr>
<td>30</td>
<td>2.959031</td>
</tr>
<tr>
<td>50</td>
<td>8.219531</td>
</tr>
<tr>
<td>75</td>
<td>18.49395</td>
</tr>
<tr>
<td>100</td>
<td>32.87812</td>
</tr>
<tr>
<td>150</td>
<td>73.97578</td>
</tr>
<tr>
<td>170</td>
<td>95.01778</td>
</tr>
<tr>
<td>174.4</td>
<td>100</td>
</tr>
<tr>
<td>175</td>
<td>100</td>
</tr>
<tr>
<td>180</td>
<td>100</td>
</tr>
</tbody>
</table>

Inter Cluster Movement starts from the distance 174.4 (shown in bold) in column 1. Hence probability of migration is 100% from 174.4.

The probability $(p)$ of inter cluster movement $(0 \leq p \leq 1)$ increases with increasing distance of the new data point. This is evident from Fig.2 with horizontal axis representing distance and the vertical axis is the probability of migration of data points between clusters.

The probability is 0, when the new data point is positioned very close to $c_k$, such that no inter cluster movement occurs. As the new data point moves farther away from $c_k$, the probability increases. At one stage, the probability is 1, when the distance $d'$ causing inter cluster movement is identified and continues to be 1, with increasing $d'$.

5. Implementation of regression analysis in predicting inter cluster movement

The prediction of inter cluster movement is carried out using regression analysis. In Table 1, the $X_i$ column (independent variable) is the distance $d$ from cluster center $c_1$ and $Y_i$ column (dependent variable) is the probability of migration based on geometric model. With this input, a regression equation $Y = A + BX$ is established and is used to predict the probability of migration induced by the new data point $(Y)$, given any value of distance from its center $(X)$. After solving $A$ and $B$, the regression equation becomes $Y = -19.3435798422621 + 0.60951798708451X$.

On applying the regression equation, when $X=152$, predicted probability of migration is 75.96162.
6. Experimental results and Analysis

Consider a data point induced at a distance 174 from \( c_1 \). The probability \( (P) \) of inter cluster movement was calculated by executing K Means Clustering algorithm for varying number of clusters. Also, a better prediction of probability \((P')\) of inter cluster migration was established from regression. The variation of experimental probability \((P)\) from the predicted probability \((P')\) is computed as relative error. Table 2 and Fig. 2 gives the result for the data point introduced at a distance 174 from cluster center \( c_1 \).

Table 2: Experimental probability and predicted probability results for the data point located at distance 174 from \( c_1 \) with increasing size of the 2 dimensional clusters

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Observed Geometric Probability ((P))</th>
<th>Predicted Probability from regression ((P'))</th>
<th>Relative Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>99.54181</td>
<td>98.128674</td>
<td>1.419641</td>
</tr>
<tr>
<td>3</td>
<td>78.75391</td>
<td>76.43655</td>
<td>2.942533</td>
</tr>
<tr>
<td>4</td>
<td>52.48694</td>
<td>54.8674</td>
<td>4.535338</td>
</tr>
<tr>
<td>5</td>
<td>47.9422</td>
<td>50.7634</td>
<td>5.884586</td>
</tr>
<tr>
<td>6</td>
<td>36.23851</td>
<td>38.89174</td>
<td>7.321576</td>
</tr>
<tr>
<td>7</td>
<td>28.94238</td>
<td>31.18463</td>
<td>7.74729</td>
</tr>
<tr>
<td>8</td>
<td>20.47262</td>
<td>22.571553</td>
<td>10.25239</td>
</tr>
<tr>
<td>9</td>
<td>14.428401</td>
<td>16.03273</td>
<td>11.11924</td>
</tr>
<tr>
<td>10</td>
<td>10.1761</td>
<td>11.6511</td>
<td>14.49475</td>
</tr>
</tbody>
</table>

From Figure 3, it is obvious from the results that with the increase in size of the clusters, the relative variation in probabilities of migration \((P \text{ and } P')\) also increases.

Experiments were also performed on 2 and 3 dimensional data sets for Clusters \( C_k \) \( \forall k = 2, 3, 4... 10 \). Fig.4 shows the results. The relative variation in the probabilities of migration increases with the increase in the cluster size as evident from Fig.4.

7. Conclusion

One of the behaviors of the clustering system is the movement of points between clusters to accommodate a new entrée. A probabilistic based approach wrapped with the prediction based on regression would emerge as model to forecast this cluster migration, which in turn serves to avoid repeated re-run of clustering algorithm. Thus, this possible clustering system dynamics was viewed within the purview of geometric probability and regression. On experimental results and analysis, it was inferred that an increasing relative error is due to increasing the size of the clusters and dimensions of the data. However, the variations are with significantly acceptable lower error rate and can easily be controlled.

References


Rajee.A.M received B.E degree in Computer science and engineering from Bharathidasan University, Tamilnadu, India in 2002 and M.Tech degree in Computer engineering from Manonmaniam Sundaranar University, Tamil Nadu, India in 2004. She has got 8 years of academic experience in various engineering college. Currently she is a Full time doctoral student in computer science and engineering at Pondicherry Engineering College, Puducherry, India. Her research areas are Data clustering analysis and techniques. She has published 3 research articles in various conferences and won 1 Best Paper Award.

Sagayaraj Francis.F holds a PhD in Computer Science and Engineering and M.Tech in Computer Engineering, both from Pondicherry University. He is currently working as Professor in the Department of Computer Science and Engineering at Pondicherry Engineering College, Puducherry, India. His areas of interest include Data Analysis and Knowledge discovery, Information systems, Business Intelligence. He is currently guiding 8 PhD Scholars. He is an active life member of ISTE and International Association of Computer Science and Information Technology.
Subspace identification of wind turbine shaft vibrations measured with a piezoelectric sensor

Dominik Kujawa1, Tomasz Zawadzki2

1 Institute of Control and Computation Engineering University of Zielona Góra
ul. Podgórna 50, 65-246 Zielona Góra, Poland
dominikinf@gmail.com, dkujava@weit.uz.zgora.pl
http://www.uz.zgora.pl

2 Girne Amerikan Üniversitesi,
Universite Caddesi, P.K. 5,
Karmi Kampüsü, Karaoğlanoğlu, Girne / KKTC
tomaszzawadzki@gau.edu.tr

Abstract
The paper presents a new approach to the parameter identification of a wind turbine shaft vibrations measured with piezoelectric sensor. The system order and unknown processes matrices are determined using the wind force as the model input and wind turbine shaft vibrations as the model output. The identified model can be used to estimate the shaft vibrations and design an early damage monitoring system. The real system results have been presented.

Keywords: state space models, subspace methods, identification algorithms, parameter estimation.

1. Introduction
The increasing demand for electricity along with the diminishing world’s energy resources results in increasing electricity acquiring from alternative sources such as solar energy, rivers water flow and wind energy which is known for ages when man learned to use the wind to run various types of devices. Wind power is by far the fastest growing and most common used technology of electrical energy production from renewable sources.

The contemporary designs of wind turbines are technically advanced high-efficiency devices. In order to maximize the efficiency, more efficient aerodynamic profiles are developed and the latest technologies for bearing and shaft materials are applied. Currently, the efficiency of wind turbines is similar to the efficiency of coal power plants and is approx. 30%, with 25% of the capacity (optimal wind conditions). The rapid development of technology allows to build more efficient wind turbines with higher power, but it involves an increase in the size of the structure, and hence higher costs of the construction and exploitation [1], [2].

Large structures are exposed to different kinds of failures, and downtime, in turn, is caused by their removal would require substantial financial costs. Removing and replacing defective parts at heights of tens of meters is a very complex operation. It becomes even more complicated when the workplace of turbines are open waters of the sea [3], [4].

The wind turbine shaft, into which rotor blades are attached, is particularly vulnerable to the damage. Subjected to a torsion force, a shearing stress is produced in the shaft. Due to a high price of materials used in the manufacture of a shaft, they show high mechanical strength properties.

The damage detection method proposed in the paper involves the measurement and analysis of dynamic real time parameters of the shaft (during normal operation of the plant), enabling an early detection of damage or providing with the information about the possibility of such damage in the future. The diagnostic system in newly designed structures is often an integral part of the structure of wind turbine. The monitoring of the particular parameters allows one to create the conditions to guarantee trouble-free operation for a predetermined time, or even extend the life-cycle of the part (alignment, etc.). In the event of damage, a quick reaction minimizes the impact of damage to other components, respectively. The development and implementation of such systems is of vital interest to the leading companies in this field. The use of the study results will allow to significantly reduce operating costs through optimal planning technical maintenance, increasing safety and reducing environmental hazards that may arise during the removal of the consequences of a possible failure simultaneously [5].

Every electrical device in use is subjected to various factors that cause irreversible processes causing changes in the object and the gradual deterioration of its performance features. From the user's perspective, the important part is the object state assessment without interfering with its structural properties and surface. This
assessment provides non-destructive testing means which can detect damage, thus preventing a major failure. These methods include, among others, ultrasound resonant method, acoustic emission, infrared techniques, radiography. The used techniques are based on signal processing methods such as wavelet transforms, neural networks and genetic algorithms [6], [7].

The new approach proposed in this paper uses subspace identification algorithm and to build the system model, the input-output signals are acquired from a piezoelectric sensor and a wind force sensor. Signal acquisition based on a group of vibration methods analyze dynamic parameters of the structure. At the moment of damage of property structures, factors such as: stiffness, mass, damping change as well. Thus, the modal parameters are changed (natural frequency, damping modal). Based on the input-output data analysis, changes in the object can be detected [4].

Vibratory methods are mainly based on the characteristics of natural object vibrations, for example: natural frequencies or vibration characters. The damage location and its size can be determined by finding the difference between the dynamic characteristics of damaged and undamaged state. In many studies, the natural frequency of the structure is used to show damage to the structure. The advantage of this method is the ease and accuracy of measurement. Changing the natural frequency of the structure indicates occurrence of the damage [8].

Prediction error identification algorithm and the N4SID algorithm can be used to determine the vibration model of an undamaged shaft of a wind turbine. Subspace algorithms comprise of two steps [9], [10], [11], [12], [13]. In the first step, a weighted projection of the row space of data Hankel matrices is performed and the system order and the extended observability matrix are obtained straightforwardly from the input-output signals. The second step, in turn, determines the unidentified system matrices either by means of determining the state sequences and combining them with the input-output signals [14], [15] or determining the matrices $A$ and $C$ directly from the extended observability matrix and using them to determine the remaining system matrices [16], [17]. Subspace methods do not require any canonical parameterization. They are an interesting alternative to the well-known prediction error methods due to a simple and general parameterization in the Multiple-Input Multiple-Output (MIMO) case. No nonlinear optimization is necessary and reliable state-space models for MIMO dynamic systems are derived directly from the input-output data [18]. Moreover, the computational complexity of subspace methods is modest in comparison to the well-known prediction error methods [19], [20].

The subspace identification approach proposed in the paper uses appropriate Hankel matrices built from the measured input-output signals that are susceptible to vibration frequencies and the wind force. To detect a fault, the algorithm analyzes the values of the actual system outputs and compares them with the corresponding behavior of the model built using the input-output data acquired before the fault occurrence.

In the deterministic-stochastic approach, the disorder of construction can be treated as an additional excitation causing the change in the system output. During signal measuring, we encounter the problem of noise, which may result from imperfections of measuring equipment. This is particularly applicable for low damage, and hence a small change in a signal which may be indistinguishable from noise. In this case, the proposed subspace identification method is appropriate as the possibility of damage can be analyzed.

The remaining of the paper is organized as follows. Section 2 presents the wind turbine experimental research. The identification problem is formulated and the proposed solution based on subspace algorithm is presented in Sections 3 – 5. Section 6 presents the identification results based on a real system input-output data. Conclusions are given in Section 7.

### 2. Experimental wind turbine measurements

The experimental wind turbine has been mounted on the roof of a dwelling house, Fig. 1.

![Wind turbine mounted on the roof of a house.](image)

The modification of the wind turbine has been made by placing a current disc in front of a generator, shown in Fig. 2. Nearby the current disc, a piezoelectric sensor measuring vibration in the vertical surface has been mounted.
MTN Vibration sensors are made of stainless steel for continuous monitoring of vibration in difficult conditions. The electronics inside the sensor is enclosed in a Faraday cage and insulated to minimize noise. The sensor has IP67 standard, two-wire 4-20 mA output, proportional to the range of the transmitter. It is equipped with a 2-pin MS, 4-pin M12 connector and braided ETFE stainless steel cable [21].

A two-channel data acquisition module designed for use with IEPE sensors is used in the system part. The device inputs are configured as AC. The current input type is indicated by the corresponding panel LED. The inputs can be switched using buttons. The device is fully powered by the USB port. Its small size and weight make the device easy to use.

The module has an additional amplification (1,10,100) for each channel indicated by a panel LED and switched using buttons. It also has an exceeding the input voltage range gauge for both channels also indicated by LEDs mounted on the panel [22].

3. Deterministic discrete processes

Consider the state-space model [20] of a discrete linear dynamical process of the following form:

\[ x(p+1) = Ax(p) + Bu(p) \]  \( (1) \)

\[ y(p) = Cx(p) + Du(p) \]  \( (2) \)

where:
- \( 0 \leq p \leq \alpha - 1 \in \mathbb{Z}_+ \) – the independent temporal variable,
- \( x(p) \in \mathbb{R}^n \) – the state vector,
- \( y(p) \in \mathbb{R}^l \) – the output vector,
- \( u(p) \in \mathbb{R}^m \) – the input vector,

\( A,B,C,D \) – matrices of appropriate dimensions.

To complete process description, it is necessary to specify the initial condition [6], [13]:

\[ x(0) = d \]  \( (3) \)

where \( d \in \mathbb{R}^n \) is a vector with known constant entries.

Define the following Hankel block matrices [2]:

\[
\begin{align*}
U_{0|2i-1} & \overset{\text{def}}{=} \begin{bmatrix} U_0 \\ U_1 \\ \vdots \\ U_{2i-1} \end{bmatrix} \\
Y_{0|2i-1} & \overset{\text{def}}{=} \begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_{2i-1} \end{bmatrix}
\end{align*}
\]

where

\[ U_k = [u(k)u(k+1)\cdots u(k+j-1)] \]
\[ Y_k = [y(r)y(r+1)\cdots y(r+j-1)] \]

Define also the Hankel block matrices:

\[
\begin{align*}
\frac{U_{0|i-1}}{U_{i|2i-1}} & = U_p \\
\frac{U_{0|i-1}}{U_{i+1|2i-1}} & = U'_p \\
\frac{Y_{0|i-1}}{Y_{i+1|2i-1}} & = Y_p \\
\frac{Y_{0|i-1}}{Y_{i+1|2i-1}} & = Y'_p 
\end{align*}
\]

The number of block rows \( i \) should be larger than the maximum order of the system and to use all data samples, the number of column should be equal to \( \alpha - 2i + 1 \) [13].

Define block Hankel matrices \( W_p \) and \( W'_p \) consisting of \( Y_p, U_p \) and \( Y'_p, U'_p \), respectively:

\[
W_{0|i-1} = \begin{bmatrix} U_{0|i-1} \\ Y_{0|i-1} \end{bmatrix} = \begin{bmatrix} U_p \\ Y_p \end{bmatrix}
\]

\[
W'_p = \begin{bmatrix} U'_p \\ Y'_p \end{bmatrix}
\]

The state-sequence matrix \( X_i \) is defined as:
\[ X_i = [x(i) \ldots x(i+j-1)] \]

Define the extended observability matrix \( \Gamma_i \) and the reversed extended controllability matrix \( \Delta_i \):

\[
\Gamma_i = \begin{bmatrix}
C \\
CA \\
CA^2 \\
\vdots \\
CA^{i-1}
\end{bmatrix}
\]

\[
\Delta_i = [A^{-1}[B] \ldots A[B][B]]
\]

Assume also that the pair \( \{A,C\} \) is observable and the pair \( \{A,B\} \) is controllable [14], [13]. Finally, define the lower block triangular Toeplitz matrix \( H_i \):

\[
H_i = \begin{bmatrix}
[D] & 0 & \cdots & 0 \\
C[B] & [D] & \cdots & 0 \\
CA[B] & C[B] & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
CA^{i-2}[B] & CA^{i-3}[B] & \cdots & [D]
\end{bmatrix}
\]

4. Identification problem

Given \( \alpha \) measurements of the input \( u(p) \) and the outputs \( y(p) \) generated by the process (1) – (2) determine its order and the matrices \( A,B,C,D \) up to within a similarity transformation.

Following Theorem 1 [12], the state-space model (1) – (2) can be written in a matrix form:

\[
\begin{align*}
Y_p &= \Gamma_i X_p + H_i U_p, \\
Y_f &= \Gamma_i X_f + H_i U_f, \\
X_f &= A^T X_p + \Delta U_p.
\end{align*}
\]

5. Stochastic discrete processes

Consider the state-space model of a discrete linear dynamical process of the following form:

\[
\begin{align*}
x(p+1) &= Ax(p) + Bu(p) + w(p) \\
y(p) &= Cx(p) + Du(p) + v(p)
\end{align*}
\]

where the covariance matrix of the zero mean white vector sequences \( w(p) \) and \( v(p) \) is

\[
E\left[ \begin{bmatrix} u(k) \\ v(k) \end{bmatrix} \right] = \begin{bmatrix} Q & S \\ ST & R \end{bmatrix} \delta_k (7)
\]

and \( \delta_k \) denotes the discrete Kronecker delta.

The identification problem is: given \( \alpha \) measurements of the input \( u(p) \) and the output \( y(p) \), generated by (5) and (6), determine the system order and the matrices \( A,B,C,D \) up to a similarity transformation, and the covariance matrices \( Q,S,R \).

We assume that \( u(p) \) and \( y(p) \) are uncorrelated with \( w(p) \) and \( v(p) \), \( u(p) \) and \( y(p) \) are persistently exciting of order \( 2\alpha \), and \( w(p) \) and \( v(p) \) are not identically zero.

Based on Theorem 12 [12], the combined deterministic-stochastic Algorithm 1 or its robust version can be applied to determine the process order and the unknown matrices \( A,B,C,D,Q,S,R \). The combined deterministic-stochastic Algorithm 1 consists of the following steps [12]:

1) Calculate the oblique projection

\[
O_i = Y_f / U_f \ W_p
\]

2) Calculate the singular value decomposition

\[
W_i O_i W_2 = U_i U_2 \begin{bmatrix} S_i & 0 \\ 0 & V_i^T \end{bmatrix} = U_i S_i V_i^T
\]

3) Find the order of the process (6) – (7) by the inspection of the singular values in \( S_i \).

4) Calculate \( \Gamma_i \) from

\[
\Gamma_i = W_i^{-1} U_i S_i^{-2}
\]

5) Solve the following set of linear equations in a least squares sense for \( A,C \) and \( K \):

\[
\begin{bmatrix} \Gamma_{i-1} Z_{i+1} \\ Y_{ii} \end{bmatrix} = \begin{bmatrix} A^T Z_i + K U_f + [\rho_w] \\ C \end{bmatrix}
\]

where \( Z_i \) and \( Z_{i+1} \) are the orthogonal projections

\[
Z_i = Y_f / \begin{bmatrix} W_p \\ U_f \end{bmatrix}
\]

\[
Z_{i+1} = Y_f / \begin{bmatrix} W_p^T \\ U_f^T \end{bmatrix}
\]
and \( \dagger \) denotes the Moore-Penrose pseudo-inverse of the matrix.

6) Using the least squares method, determine \( B \) and \( D \) from the following over-determined set of equations

\[
\begin{bmatrix}
K_{11} \\
\vdots \\
K_{1j} \\
K_{21} \\
\vdots \\
K_{2j}
\end{bmatrix} = N \begin{bmatrix} D \\ B \end{bmatrix}
\]  \( (14) \)

where

\[
N = \begin{bmatrix}
-\lambda_{ij3} & M_{1} - \lambda_{ij1} & \cdots & M_{i-2} - \lambda_{ij1} & M_{i-1} - \lambda_{ij1} \\
M_{1} - \lambda_{ij2} & M_{2} - \lambda_{ij2} & \cdots & M_{i-2} - \lambda_{ij2} & M_{i-1} - \lambda_{ij2} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
M_{i-1} - \lambda_{ij1} & 0 & \cdots & 0 & 0 \\
I_{i} - \lambda_{ij1} & -\lambda_{ij2} & \cdots & -\lambda_{ij2} & -\lambda_{ij1} \\
-\lambda_{ij3} & -\lambda_{ij3} & \cdots & -\lambda_{ij3} & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
-\lambda_{ij1} & 0 & \cdots & -\lambda_{ij1} & 0
\end{bmatrix}
\]

\[
\times \begin{bmatrix} I_{i} \\ 0 \\ 0 \end{bmatrix}
\]  \( (15) \)

with

\[
\lambda = \begin{bmatrix} A \\ C \end{bmatrix} \Gamma_{i}^\dagger = \begin{bmatrix} \lambda_{ij1} & \lambda_{ij2} & \cdots & \lambda_{ij1} \\
\lambda_{ij2} & \lambda_{ij2} & \cdots & \lambda_{ij2} \end{bmatrix}
\]  \( (16) \)

\[
M = \Gamma_{i}^\dagger = \begin{bmatrix} M_{1} & M_{2} & \cdots & M_{i-1} \\
K_{11} & K_{12} & \cdots & K_{1i} \\
K_{21} & K_{22} & \cdots & K_{2i} 
\end{bmatrix}
\]  \( (17) \)

\[
K = \begin{bmatrix} K_{11} \\
K_{12} \\
\vdots \\
K_{21} \\
K_{22} \\
K_{2i}
\end{bmatrix}
\]  \( (18) \)

7) Determine \( Q, S \) and \( R \) from the residuals \( \rho_{w} \) and \( \rho_{v} \)

\[
\begin{bmatrix} Q \\ S^T \\ R \end{bmatrix} = E \begin{bmatrix} \rho_{w} \\ \rho_{v} \end{bmatrix} \begin{bmatrix} \rho_{w}^T \\ \rho_{v}^T \end{bmatrix}
\]  \( (19) \)

6. Identification results

The combined subspace identification algorithm was tested with the experimental input-output data set. The data set was created in the following way: for 1280 values, each value consists of 20 averaged trials where each trail contains the wind force and vibrations values. The test was performed in the following way: averaged values of vibrations and the wind force, observed within a minute for a properly operating wind turbine, for the test time were included to the model construction. The observed piezoelectric sensor values are shown in Fig. 3 and the change of wind force is shown in Fig. 4. The data acquisition time is 1280 minutes.

Fig. 5 shows the responses of the system and the second order model; fit to the actual values is 66%. Fig. 6 shows the 4th order model response which fit to the actual values is 89%.
Innovation Strategies, Measure 8.2 Transfer of knowledge, the input-output data set was created using the wind, the stronger are the vibrations. For the subspace input and output values averaged in one minute windows.

The identification experiments are performed on the basis of real measurement input-output data. To identify the system model, the combined deterministic-stochastic identification algorithm was applied with the wind force as the input and the vibration rate as the output.

**REFERENCES**


**ACKNOWLEDGMENT**

The author is a scholar within Sub-measure 8.2.2 Regional Innovation Strategies, Measure 8.2 Transfer of knowledge, Priority VIII Regional human resources for the economy Human Capital Operational Programme co-financed by the European Social Fund and the State Budget of Poland.

**7. Conclusions**

The paper presents a new approach to the subspace identification of a wind turbine vibrations. Wind force has an influence on the vibration of the shaft, the stronger is the wind, the stronger are the vibrations. For the subspace identification, the input-output data set was created using input and output values averaged in one minute windows. In the future, models of various types of damage will be developed, e.g. rotor blades or a generator. The identification experiments are performed on the basis of real measurement input-output data. To identify the system model, the combined deterministic-stochastic identification algorithm was applied with the wind force as the input and the vibration rate as the output.
A Hybrid Approach to Privacy Preserving in Association Rules Mining

Narges Jamshidian Ghalehsefidi¹, Mohammad Naderi Dehkordi²

¹²Department of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran
¹jamshidian_n@sco iaun.ac.ir, ²naderi@iaun.ac.ir

Abstract

Nowadays, data mining is a useful, yet dangerous technology through which useful information and the relationships between items in a database are detected. Today, companies and users need to share information with others for their progress and they should somehow manage this information sharing for preserving sensitive information. Privacy preserving in data mining was introduced for managing information sharing. This paper presents a hybrid algorithm with distortion technique with both support-based and confidence-based approaches for privacy preserving. The proposed algorithm tries to maintain useful association rules and hide sensitive rules from the perspective of the database owner. It also has no limit on the number of items in the left-hand side and the right-hand side of rules. This paper also compares the proposed algorithm with MDSRRC algorithm and 1.b algorithm. The proposed algorithm has less lost rules compared with the MDSRRC and 1.b algorithms and its CPU usage is less then.

Keywords: privacy preserving, hiding sensitive rules, helpful association rules.

1. Introduction

The data mining technique aims to detect useful rules and relationships of database items for which standard algorithms are used such as Apriori and Eclat. The detected rules are divided into two groups: sensitive rules and non-sensitive rules. Sensitive rules are those rules the database owner is trying to hide using privacy preserving algorithms in data mining, and non-sensitive rules are useful rules the database owner wants to share with others. Of course, doing anything has its own costs. The cost the database owner pays for hiding sensitive rules is the loss of non-sensitive rules plus other costs we called them side effects due to hiding sensitive rules. These side effects include loss of non-sensitive rules, creating ghost rules, hiding failure, dissimilarity, runtime, etc. The algorithms presented in the context of privacy preserving have tried to reduce these side effects. The proposed algorithm aims to reduce lost rules and reduce hiding failure to zero. Generally there are two approaches for hiding sensitive rules [1]: support-based approach and confidence-based approach. The support-based approach aims to reduce support of the sensitive rule by reducing support of one of element sets composing the sensitive rule. The confidence-based approach aims is reducing confidence of the sensitive rule through increasing support of the consequent of the sensitive rule. This paper uses both approaches for hiding sensitive rules. The proposed algorithm selects either approach to sanitize by calculating the support of left-hand side and right-hand side elements of the sensitive rule. In this algorithm, selecting the item and the transaction to sanitize has a large impact on reducing side effects. This paper compares the proposed algorithm with MDSRRC algorithm on the Chess dataset.

The article is organized as follows. Section 2 describes the framework of association rule. Section 3 reviews the related works. Section 4 explains the proposed algorithm, terms and steps. Section 5 compares and evaluates the proposed algorithm with MDSRRC and 1.b. Finally, conclusions are presented in Section 6.

2. Framework of association rules

Rule extraction in data mining is performed by the level of support and confidence of the rule. The issue of extracting association rule was introduced by [2]. Suppose $I=\{i_1,i_2,\ldots,i_m\}$ is a set of elements and the database $D=\{T_1,\ldots,T_n\}$ is a set of transactions. Each transaction $T\in D$ contains a subset of $I$. The general framework of association rules is $X\rightarrow Y$. If $X$ and $Y$ are subsets of $I$ and if $x\cap y=\emptyset$, then $X$ is called the antecedent or LHS of the rule and $Y$ is called the consequent or RHS.

The support of the rule $X\rightarrow Y$ is defined by calculating the ratio of simultaneous repetition frequencies of $X$ and $Y$ in transactions to the total number of transactions in the database. The support of the rule is calculated by Eq. (1).

$$\text{support}(X\rightarrow Y) = \frac{|X \cup Y|}{|D|}$$

(1)

The confidence of the rule $X\rightarrow Y$ is defined by calculating the ratio of simultaneous repetition frequencies of $X$ and $Y$ in
transactions to the number of repetition of X alone in transactions of the database. The confidence of the rule is calculated by Eq. (2).

\[
\text{confidence (X→Y)} = \frac{|X \cup Y|}{|X|}
\]

(2)

Minimum Support Threshold (MST) and Minimum Confidence Threshold (MCT) criteria are used to extract useful rules from the database. If Support \( (x→y) \geq MST \) and Confidence \( (x→y) \geq MCT \), the \( x→y \) rule becomes important and is extracted from the database at the time of data mining.

3. Related Works

Wang et al. presented two algorithms for hiding association rules. The first algorithm, ISL, reduces the rule confidence by increasing support of the left-hand side element set of the sensitive rule. This algorithm has high hiding failure and new-rule creation. The second algorithm, DSR, reduces confidence by reducing support of the right-hand side element set of the rule. The failure of this algorithm is close to zero but many non-sensitive rules are lost [3] [4].

Modi et al. introduced an algorithm called DSRRC which uses clustering of right-hand side common items for hiding. Its disadvantages are as follows: it only hides rules with an element at their right-hand side, it is dependent on the ordering of transactions, gives a different result with reordering transactions in the database, needs sorting of the database after each item is deleted and is not suitable for large databases. Lost rules in this algorithm are high [5].

Komal Shah et al. presented an algorithm called ADSRRC to modify DSRRC. This algorithm also hides the rules with a single RHS, and it sorts transactions only once. In addition, this paper proposed a new algorithm called RRLR which hides rules with a single LHS. It reduces both support and confidence for hiding sensitive rules [6].

To overcome limitations in the left-hand side and right-hand side items, Domadiya et al. proposed an algorithm called MDSRRRC. It selects the best item for deletion based on the number of its repetition in the right-hand side of the sensitive rule and its support. This algorithm has less side effects compared to DSRRC. It fails in certain circumstances [7].

Kumar Jain et al. combined the two algorithms ISL and DSR and stated the main goal as reducing the number of database changes and reducing the time for hiding sensitive rules [8].

Vijayarani et al. presented a heuristic algorithm called ABC. In this algorithm, transaction selection is performed randomly according to the behavior of honey bees for finding food. This algorithm uses the support-based approach [9].

Oliveria et al. presented two algorithms called Round Robin and Random. The essence of both algorithms is item selection to sanitize which is done randomly and intermittently [10].

Duraiswamy et al. proposed an algorithm called SRH. It reduces complexity, time and memory by calculating the number of transactions required for hiding sensitive rules [11].

Menon et al. proposed an algorithm with exact approach called Integer programming and also two strategies: Blanket and Intelligent. This algorithm has the best level of accuracy [12].

Verykios et al. proposed two algorithms called WSDA and BA. WSDA hides sensitive rules using the distortion techniques, and BA does the same using the blocking technique [13].

Amiri proposed three algorithms called Aggregate, Disaggregate and Hybrid, which hide sensitive rules using the support-based approach [14].

Dasseni et al. generalized the hiding problem to a combination of sensitive rules hiding and sensitive itemset hiding. Algorithm 1b selects the best subset from the itemset on the right side of the sensitive rule and for hiding the sensitive rule selects the first item as the victim item and eliminates the latter from those transactions fully supporting the sensitive rule. The advantages of this algorithm are in reducing hiding failure and relatively proper CPU usage [15].

4. The proposed algorithm

The proposed approach uses the distortion technique for hiding association rules with both confidence-based and support-based approaches. This algorithm has two main goals:

1. Reduction of non-sensitive lost rules due to hiding sensitive rules.
2. Reduced CPU usage.

We first introduce the terminology used in the proposed algorithm and then describe its steps.

Sensitive item and item sensitivity: Items involved in sensitive rules are called sensitive items, and the number of their repetition in sensitive rules is called sensitive items.

Degree of transaction collision: The number of sensitive rules in the transaction. The transaction in fact contains all items involved in the sensitive rule.

4.1 The proposed algorithm steps

The essence of the proposed algorithm is the selection of sanitizing operation according to the amount of LHS and RHS support of the sensitive rule. Then the sensitive item and the suitable transaction are selected for the sanitizing operation. The sensitive item is selected considering the amount of support and sensitivity of that item which leads to selecting the suitable item for the sanitizing operation.

Before the sanitizing operation, the proposed algorithm first obtains the number of transactions required for hiding sensitive
rules using the equation presented in Section [3]. Eq. (3) shows how to calculate mincount.

MSC x→y = count (xUy) - |D| * min support + 1.
MCCC x→y = count (xUy) - [count(x) * min conf] + 1.
MPCC x→y = [(count (xUy) - count(x) * min conf) / (1-min conf)] + 1.
MCC = minimum (MCCC, MPCC).
mincount = minimum (MSC, MCC).

**PSEUDO CODE FOR Proposed Algorithm**

Input: Source Database D, MCT, MST, Sensitive rule (R_H)
Output: The Sanitized database D

1. Find sensitivity of each item ∈ R_H set I_S
2. Find support of each item ∈ I_S in D
3. Find conflict T∈D set T_S
4. Sort R_H by decreasing order of their support
5. Hiding
6. While(all the sensitive rule hidden≠true) {
   a. Foreach r in R_H do{
      i. If support r_{LHS} >= Support r_{RHS} {
         1. Sort I_S by sensitive item decreasing, support increasing
         2. Select Victim item where RHS contains it
         3. Mincount =mincount(r)
         4. Sort T_S by conflict decreasing, length increasing
         5. Foreach t in T_S do{
            a. If(itemset xyz ∈ t) {
               i. Remove itemselected
               ii. If Mincount=0 then break
            }
         }  
         6.  }  
      ii. Else{
         1. Sort I_S by support increasing, sensitive item decreasing
         2. Select sensitive item where LHS contains it
         3. Mincount =mincount(r)
         4. Sort T_S by conflict decreasing, length increasing
         5. Foreach t in T_S do{
            a. If(itemset xyz ∈ t) {
               i. Remove itemselected
               ii. If Mincount=0 then break
            }
         } 
         6. } 
      } 
   } 
   b. Start Update support& confidence
   7. }

5. Comparison and evaluation

We implemented the proposed algorithm with the known MDSRRRC algorithm and l.b algorithm on the Sony F115FM system with CPU Core i7, memory 6GB, HDD 500GB, Windows 7 OS with the C# programming language. For comparison and evaluation purposes, we performed various tests with different rules on the Chess dataset. Figure 1 shows the graph of non-sensitive lost rules and figure 2 shows the graph of CPU usage in the tests conducted on the Chess dataset.

![Fig. 1 Lost rule on the Chess dataset](image-url)
6. Conclusions

This paper proposed a hybrid algorithm with two support-based and confidence-based approaches. The proposed algorithm offers better results in actual compressed databases compared with dummy or actual non-compressed databases. To solve this problem, we can simply change the sorting of sensitive items with dummy or actual non-compressed databases. To solve this problem, we can simply change the sorting of sensitive items with dummy or actual non-compressed databases. In the MDSRRC algorithm, there is a possibility of failure in certain circumstances due to uncontrolled hidden rules. We address this problem by controlling sensitive rules even after being hidden. The hiding failure in the proposed algorithm is zero. The proposed algorithm maintains more non-sensitive rules than MDSRRC and 1.b algorithm.

References

Narges Jamshidian Ghalehsefid was born in Isfahan, Iran in 1988. She received a B.S. degree in Computer Engineering from Najafabad Branch, Islamic Azad University, Najafabad, Iran, and is currently an M.S. student in Computer Engineering (major: Software) at this university. Her field of research is Preserving Privacy in Data Mining.
A Game Theoretical Interest Forwarding for Cached Data in Content-Centric Networking

Chengming Li¹ and Koji Okamura²

¹,² Department of Advanced IT, Graduate School of ISEE, Kyushu University
Hakozaki 6-10-1, Higashi-ku, Fukuoka, Japan
Chengming.li.dut@gmail.com

² Research Institute for Information Technology, Kyushu University
Hakozaki 6-10-1, Higashi-ku, Fukuoka, Japan
oka@ec.kyushu-u.ac.jp

Abstract
Content-Centric Networking (CCN) has recently emerged as a clean slate approach to rethink Internet foundations, which changes from host-centric communication model to content-centric. It is common that the current router does not have all the information of cached data in network, because of the huge naming space and volatility of Content Store in each router. In this paper, we argue that it is necessary to supplement CCN with mechanisms to make multiple Interests forwarding for cached data. Our goal is to maximize the residual capacity in the network so that users can get the maximum payoff in a definite network situation. We proposed a game theoretical Interest Forwarding Decisions to analysis the properties of user behavior. Evaluation results prove that our proposals improve user’s payoff in the light load case for content-centric networking.

Keywords: Game Theory, Nash Equilibrium, Content-Centric Networking, Interest Forwarding Decisions.

1. Introduction
The architecture of today’s Internet is originally designed as a communication model that is a conversation between exactly two machines. However, content traffic has been increasingly prevalent in the Internet. Some video content providers (CPs, e.g., YouTube and Hulu) have even begun to provide high-definition video streaming services. As demand for highly scalable and efficient distribution of content increases, the TCP/IP architecture may reveal its inefficiency in delivering time-sensitive multimedia traffic [3]. It now mostly serves content-centric applications, e.g., Content Distribution Networks (CDNs) [11] and P2P. The Internet architecture has evolved substantially from host-centric communication model to content-centric model.

There are a number of proposed architectures for Information-Centric Networking (ICN) including the Publish Subscribe Internet Routing Paradigm (PSIRP) [8], the Network of Information (NetInf) from the Design for the Future Internet (4WARD) [4], the Cache-and-Forward Network Architecture [5], the Data Oriented Network Architecture (DONA) [7], and the Content Centric Networking (CCN) [6].

Content-Centric Networking (CCN) (It is also called Named Data Networking) [18] is designed inherently to focus on content distribution rather than host-to-host connectivity. CCN retrieves a content object by its name, instead of its storage location in order to address IP network’s limitations in supporting content distribution. This change, decoupling content from hosts at the network layer, has several attractive advantages, such as network load reduction, low dissemination latency and energy efficiency.

It is a challenge that how to efficiently utilize the cached data. In some cases, the content objects are so many that the CS cannot efficiently manage them, which may result in poor caching performance. Forwarding Information Base (FIB) of routers cannot contain all the content as the huge naming space; and as the content cached in Content Store of routers is changing frequently, it is very difficult to update the FIB in time for all content objects in the network. Thus, it is a problem that how to search the cached data efficiently.

Forwarding strategy is a key component in CCN nodes that makes them more powerful than their IP counterparts. Routing of IP network is to calculate a single shortest path for each pair of source node and destination node. The forwarding strategy layer in a CCN node can dynamically select multiple interfaces from the FIB to forward a same Interest packet. Single shortest path can be a candidate forwarding strategy for CCN. However, it cannot perform well as it runs in end to end communication network. In end to end communication network the destination node is definite, but in CCN a content object can have multiple destination nodes (it can be a router) by the form of replicas.
It is necessary to supplement CCN with mechanisms making the Interest forwarding decisions. In the case of sufficient network resources, delivering the Interest packet to multiple interfaces derived for FIB can achieve following advantages:

- The real-time decision enables nodes to fully utilize their rich connectivity and get the best users’ payoff;
- It defends against route hijacking attacks (if no data returns over a particular interface for a particular name, that interface may not lead to a valid path for that name);
- It enhances the network instability (frequent oscillation of paths) while maintaining good data delivery performance.

In this paper, we proposed a game theoretical Interest multiple forwarding decisions method to maximize the users’ payoff and network’s payoff.

The rest of this paper is organized as follows. Background and related work are given in Section 2. Section 3 presents the non-cooperative game analysis for Interest multiple forwarding problems. Section 4 presents simulation setting and simulation results. Section 5 concludes the paper.

2. Background

2.1 Content-Centric Networking

CCN design assumes a name may be viewed as a hierarchical structure of byte strings, e.g., a movie produced by Youtube may have the name “/Youtube/movies/Example.rmvb”. A node in CCN contains three data structures: the Content Store (CS), the Pending Interest Table (PIT), and the Forwarding Information Base (FIB) [18]. The structure of a CCN FIB is similar to that of an IP FIB except that CCN allows a match to multiple outgoing links. In addition, a longest-prefix match in FIB uses a content name instead of an IP address.

Communication in CCN is driven by the receiving end, i.e., the data consumer. To receive data, a consumer sends out an Interest packet which carries a name that identifies the desired data. When the Interest Packet arrives at a CCN router, the node consults the CS, PIT and FIB in sequence. The router first checks whether the data requested have already been cached in the node’s Content Store (CS) which is used to store the coming data packet by a cache replacement policy. If there is no matched data, the router will check whether the PIT has included the same Interest. In PIT, each entry contains the name of Interest and a set of interfaces from which the Interest packets have been received. If the PIT already has contained the same Interest, then the node adds the Interest coming interface to the corresponding entry of PIT. Finally, the node remembers the interface from which the request comes, and then forwards the Interest packet by looking up the name in its FIB, which is populated by a name-based routing protocol.

Once the Interest reaches a node which contains the requested data, a Data packet, which carries both the name and the content of the data, is sent back together with a signature signed by the producer’s key. This Data packet trace in the reverse path created by the Interest packet back to the consumer.

2.2 Game Theory

John von Neumann and Oskar Morgenstern established game theory as a separate field of science when they published their book in 1944[17]. Since then great strides have been made in this area, mainly in the field of economics and biology. However, game theory can also be applied to many fields of science, where decision makers have conflicting interests. Thus, it comes as no surprise to read papers related to networking that adopt game theoretical concepts to analyze a protocol’s performance or propose a solution that corresponds to a Nash Equilibrium (NE) set of strategies [2][12].

Game theory could be defined as "the study of mathematical models of conflict and cooperation between intelligent rational decision makers" [9].

A game consists of a principal and a finite set of players \( N = \{ 1, 2, \ldots, N \} \), each of which selects a strategy \( \pi_i \in X_i \) with the objective of maximizing his utility \( u_i \). The utility function \( u_i : X \rightarrow R \) represents each player’s sensitivity to everyone’s actions. People or entities (decision makers in general) who play the game are called the players.

A strategy for a player is a complete plan of actions in all possible situations in the game. The players try to act selfishly to maximize their consequences according to their preferences. The set of player \( i \)'s possible actions is called the action space \( X_i \) of player \( i \).

Two types of games are distinguished: one is non-cooperative games in which each player selects strategies without coordination with others. The other is cooperative games in which the players cooperatively try to come to an agreement, and the players have a choice to bargain with each other so that they can gain maximum benefit, which
3. Design

In this section, we firstly analysis the problem, then construct a game theoretical model to solve it. At last, we proposed Potential Heuristic Allocation for System.

3.1 Problem Description

Forwarding strategy layer, a key component of CCN nodes, make them more powerful than their IP counterparts. Routing of IP network is just to calculate a single shortest path for each pair of source node and destination node. In contrast CCN inherently supports multiple same Interests forwarding simultaneously. The forwarding strategy layer in a CCN node can dynamically select multiple interfaces from the Forwarding Information Base (FIB) to forward an Interest packet.

The simplest strategy is to send an Interest to each interface of a FIB entry in sequence. If there is no response to the Interest, then try the next interface. Single shortest path can be a candidate forwarding strategy for CCN. However, it cannot perform well as it runs in end to end communication network. In end to end communication network, the destination node is definite, but in CCN, a content object can have many destination nodes (it can be a router) by the form of replicas. Thus, sometimes the shortest path record in FIB is not real shortest path for a content object. It is very difficult to update the FIB in time for all content objects in the network because of the huge content name space, especially in chunk level.

We can also send Interests on all the interfaces at once and see which interfaces receive data first. These interfaces will be used for a period of time and their performances are monitored. If we do it for all the Interest packets, it can make the network overload and congestion easily.

A more flexible design is each FIB entry containing a program specialized to make Interest multiple forwarding decisions. In this section, we present the game theoretical Interest multiple forwarding decisions method to solve this problem. The goal of our proposals is fully utilizing the residual capacity in the network so that users can get the maximum payoff in a definite network situation.

3.2 Gaming Analysis

The hierarchical CCN naming convention described in Section 2.1 lends itself to the identification of flows. A CCN flow consists of packets bearing the same object name [19]. In a node of CCN, a set of flows \( I \) share a set of parallel paths represented by faces \( F \). Each \( F_i \in F \) has a queue length limit on how fast Interest packets can be forwarded over a face and experimented with a simple calculation of the Interest rate limit: \( |F_i| = \alpha \times C_i \div S_i \) proposed in [20]. \( |F| \) represents the maximum queue

|\( F \) represents the maximum queue length limit on how fast Interest packets can be forwarded over a face and experimented with a simple calculation of the Interest rate limit: \( |F_i| = \alpha \times C_i \div S_i \) proposed in [20]. \( |F| \) represents the maximum queue length limit on how fast Interest packets can be forwarded over a face and experimented with a simple calculation of the Interest rate limit: \( |F_i| = \alpha \times C_i \div S_i \) proposed in [20]. |\( |F| \) represents the maximum queue length limit on how fast Interest packets can be forwarded over a face and experimented with a simple calculation of the Interest rate limit: \( |F_i| = \alpha \times C_i \div S_i \) proposed in [20]. \|\( |F| \) represents the maximum queue length limit on how fast Interest packets can be forwarded over a face and experimented with a simple calculation of the Interest rate limit: \( |F_i| = \alpha \times C_i \div S_i \) proposed in [20]. ||\( |F| \) represents the maximum queue length limit on how fast Interest packets can be forwarded over a face and experimented with a simple calculation of the Interest rate limit: \( |F_i| = \alpha \times C_i \div S_i \) proposed in [20]. ||\( |F| \) represents the maximum queue length limit on how fast Interest packets can be forwarded over a face and experimented with a simple calculation of the Interest rate limit: \( |F_i| = \alpha \times C_i \div S_i \) proposed in [20].
length of face $i$ in node; $C_i$ is the upstream link capacity of face $i$; $S_i$ is an estimate of the size of the Data packets that have been received over $i$, and $\alpha$ is a configurable parameter. Here we define $X$ as the total queue length (total available resources) in the node and $X_0$ as the queue utilization caused by all background traffics. $|F|$ denotes the number of total faces in node.

$$X = \sum_{i=1}^{\mid F \mid} | F_i | \quad (1)$$

Each $I_i \in I$ aims to minimize the individual cost and maximize the utilization selfishly by deciding the multiple forwarding degree $x_i$. $F'$ is the set of faces for Interest $I_i \in I$, through which the Interest $I_i$ can reach the repository nodes with $H$ hops. We can get $F'$ from the FIB table of node. The multiple forwarding decisions problem models as $I_i$ selecting the subset $f' \subset F'$, $f' \neq \emptyset$ to get the best cost. The network model is described in Fig.1.

In our model, the game players are considered as flow $I$. Set the player $i$ using the node resources as $x_i \in X_i$. $X_i$ is a collection of node resources may be occupied by player $i$. $X_i$ is strategic space of player $i$. When only discuss faces without considering other types of node resources, $x_i$ is the multiple interfaces $f'$ used by player $i$. In our model, we define the $x_i = \{x_i \mid 1 \leq x_i \leq f' \}$ simply. For an Interest of $I$ flow, we can get the $f'$ from FIB table in CCN. In FIB table, the interfaces are sorted by the hops which present the distance from the node to repository. The face for an Interest with minimum hops has the highest priority to be selected. The Interest forwarded by the shortest path is called main Interest, correspondingly, the Interests forwarded by longer path are called replica Interest in this paper. Here, we consider the strategic space as continuously divisible to guarantee the Nash Equilibrium existence.

### 3.2.1 Payoff function

The Payoff Function of player $i$ specifies the total gains of player $i$ when it takes action $x_i$, which is a kind of variable showing the worth achieved by players using the node resources.

The general form of Payoff Function consists of two parts: $\text{Payoff} = \text{Benefit} - \text{Cost}$ [21]. Thus, the payoff function $U_i$ of player $i$ is defined as:

$$U_i(x_i, X_i) = \text{Benefit}(x_i, X_i) - \text{Cost}(x_i, X_i) \quad (2)$$

Here, $x_i$ denotes the Interest multiple forwarding degree. When there are no external controls, utility function stipulates the gain of player $i$ when it takes action $x_i$. Due to the related form of action in the game, the utility function of player $i$ is not just the function of $x_i$ but also is the function of other players.

Denote $X = (x_1, \ldots, x_n, \ldots, x_i)$ as the vector constituted by all the players’ actions, and $X_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ as the vector constituted by other players’ actions except user $i$. Then the utility obtained by player $i$ is $U_i(x_i, X_{-i})$, can also be abbreviated as $u_i(X)$. The utility function of player $i$ is the mapping from the set of action $X_i$ to the set of real number $R^+$, $u : X_i \rightarrow R^+$, which defines the preferences of players on the set of actions. For all $x, y \in X_i$, if and only if $U(x) > U(y)$, players prefer the action $x$ than the action $y$.

Here we assume that $|F'|$ is continuously divisible, and can be represented by a real number. Strategic space $X_i$ is the real axis of a non-empty closed space and is a non-empty compact convex set. It is used to guarantee the Nash Equilibrium existence. Actually, the players do the action by the rounding of $x_i$ in the simulation section.

We define the $\text{Benefit}$ function of player $i$ as:

$$\text{Benefit}(x_i, X_i) = t_m \times P \times (x_i - 1) \quad (3)$$

Here, the $t_m$ denotes unit time gain for player $i$ doing the action of sending replica Interest; $P_i$ represents the probability of a replica Interest retrieving a cached data faster than main Interest. The purpose of players who send replica Interests is to more stably retrieve the data faster.
We use $t_n \times P_i \times (x_i - 1)$ to denote the estimated benefits for players who send $x_i - 1$ replica Interests.

Cost function specifies the punishment given to players from the network when player $i$ takes action $x_i$. The Cost function is defined as:

$$Cost(x_i, X_{-i}) = t_q \times \sum_x x \times e^{(\sum x_i)(x - 1)}$$ (4)

Where the $t_q$ denotes unit time cost for queuing because of player $i$ doing the action of sending replica Interest; In this expression, the deterministic term $1 / X - (\sum x + X_0)$ represents the expected congestion delay on a link for an M/M/1 delay function [22]. We use $e^{(\sum x_i)(x - 1)}$ to express the normalized queuing time factor and adopt $\sum_x x$ to present the proportion of queuing time for player $i$.

From the network’s perspective, the nodes adopt some mechanism to transport packets efficiently and fairly. Usually the nodes use Max-Min fair queue to implement transmission fairly. We also proposed a Potential Heuristic queue method to consider efficiency and fairness in Section 3.3.

Thus, the utility function can be described as following:

$$U_i(x_i, X_{-i}) = t_n \times P_i \times (x_i - 1) - t_q \times \sum_x x \times e^{(\sum x_i)(X - 1)}$$ (5)

$U_i$ is a increasing function of $x_i$ and it is diminishing marginal returns. A higher $x_i$ does not necessarily yield better performance for player $i$. On the condition of $\sum x \leq X$, user can get an optimal $\bar{U}_i$ to meet $\frac{\partial \bar{U}_i}{\partial x_i} = 0$ through adjusting $x_i$. The unilaterally optimizing behaviors of user $i$ meet:

$$\frac{\partial \text{Benefit}_i}{\partial x_i} = \frac{\partial \text{Cost}_i}{\partial m_i}$$

(6)

Here, we assume the amount available resources of node is $X$. The resources allocation accord player’s need. We adopt a simple resource allocation method which is denoted as following:

$$x_i = \frac{|F_i|}{\sum_{j \in F_i} |F_j|} X$$ (7)

3.2.2 Nash Equilibria

A NE is a set of strategies where each player has no incentive to deviate, in other words, given the strategies of all other players, if he changes his strategy he can only decrease his utility. More specifically, if $x_i$ is an arbitrary action of player $i$ and $X_{-i}$ is the set of actions of all other players, then the action profile $x^* = (x_i^*, X_{-i}^*)$ constitutes a NE if for every player $i$, $U_i(x_i^*, X_{-i}^*) \geq U_i(x_i, X_{-i}^*)$, $\forall x_i \in X_i$, $\forall i \in [1,n]$. We set the action vector $x^* = (x_1^*, \ldots, x_n^*)$ is Nash Equilibrium, then we can get outcome: $U_i(x_i^*, X_i^*) \geq U_i(x_i, X_i^*)$, $\forall x_i \in X_i$, $\forall i \in [1,n]$.

The existence of the Nash Equilibrium [9] is constrained as following: In game $G = [n, \{x_i\}, \{U_i(\cdot)\}]$, the necessary and sufficient conditions of the existence of the Nash Equilibrium is: for all $i = 1, 2, \cdots, n$, there is: i) $X_i$ is a non-empty, compact convex set on Euclidean space; ii) $U_i(x)$ is continuous in the $x_i$, and is quasi-concave function of $x_i$.

The optimal payoff of player $i$ is recorded as $\bar{U}_i$. $\bar{U}_i$ can be assumed as increasing functions of $x_i$ (Allocated more faces, get the greater utility), and meet diminishing marginal returns (the speed of utility increasing reduces with the increase of the forwarding degree $x_i$):

$$\frac{\partial \bar{U}_i(x_i)}{\partial x_i} > 0, \quad \frac{\partial^2 \bar{U}_i(x_i)}{\partial x_i^2} < 0$$

(8)

Max $\sum_{i=1}^{n} \bar{U}_i(x_i)$ s.t. $\sum_{i=1}^{n} x_i \leq X$ (9)

The solution of our model can be represented as Eq. (9) to solve the maximum value of payoff of all players. Using Lagrange Method of Multiplier for solving, suppose a Lagrangian function $L(x_i, x_j, \cdots x_n)$ where exits:

$$L = \sum_{i=1}^{n} \bar{U}_i(x_i) + \lambda \left(X - \sum_{i=1}^{n} x_i \right)$$ (10)

In which $\lambda$ is a specific unknown constant. The optimal solution should satisfy the condition that the partial derivatives that $L$ for all unknowns is:

$$\frac{\partial L}{\partial x_i} = \frac{d \bar{U}_i}{dx_i} - \lambda = 0, i = 1, 2, \cdots, n$$ (11)

That is:

$$\frac{d \bar{U}_1}{dx_1} = \cdots = \frac{d \bar{U}_i}{dx_i} = \cdots = \frac{d \bar{U}_n}{dx_n}$$ (12)
From Eq. (5), we see that the utility function \( U_i \) is concave function. Thus, Eq. (12) has unique solution. This solution is the best Interest forwarding decisions.

3.3 Potential Heuristic Allocation for System

In our proposed model, when the node receives a set of Interest flow \( I \) with corresponding multiple forwarding decision \( x_i \), how to allocate the queue resources for each player \( I_i \in I \) is a key issue. The allocation according to user's need and fairness allocation method are not the best method because that they do not consider the system utility.

Usually, there is no global objective function of networking outcome in our proposed model or other similar models [16]. In order to improve the efficiency of whole networking, we proposed a Potential Heuristic Allocation (PHA) method using for our model. We define the global objectives of networking are 1) considering fairness of each player, 2) maximizing the player's utility and 3) improving the global networking cache hit rate.

The key idea of PHA method is that the Interest \( i \) with more potential hit has higher priority to allocate resource. For this purpose, we redesign the FIB table to record some metrics used to calculate the potential values. We add a column into FIB table named 'hits' which represents the number of hits for a Content ID by interface \( f_{j}^{ID} \). An example of FIB table is illustrated in Fig. 2.

In our model, \( k_i^j \) denotes the hits of player \( i \) through the face \( j \). The corresponding potential value \( \rho_i^j \) is defined as following:

\[
\rho_i^j = \frac{k_i^j}{\sum_{r=1}^{n} k_r^j} \quad (13)
\]

The potential value \( \rho_i^j \) implies the probability of hit for Interest \( i \) through interface \( j \). The interface list for Interest \( I \) is sorted by the value \( \rho \). Thus, \( \rho_i \) has the highest priority for player \( i \).

In PHA method, 1) the node sorts \( x_i \). The player \( I_i \) with smallest \( x_i \) has the highest priority. 2) The node sort the \( \rho_i^j \) for the players who have same \( x_i \) value. The sorting algorithm compares two \( \rho \) by priority firstly. If the priority is same, then compare the real value of two \( \rho \). 3) The node allocates the resources for each \( \rho_i^j \) by the sorted sequence until the capacity of each interface reaches the threshold \( |F_j| \) or all \( \rho_j \) has been allocated.

An example is described in Table 1. The actions of all players are \( x_1 = x_2 = x_3 = x_4 = 4 \). In our PHA method, we use the priority queue to represent the fairness. This parameter keeps that network resources can be allocated to each user fairly. The parameter \( \rho \) denotes the network utility. Under the premise of ensuring fair, we consider the network efficiency. We allocate the network resources to the players who have more probability to get cached data.

<table>
<thead>
<tr>
<th>Content Name</th>
<th>Interfaces</th>
<th>Fits</th>
<th>Potential</th>
</tr>
</thead>
<tbody>
<tr>
<td>Youtube</td>
<td>A</td>
<td>k_1</td>
<td>( \rho_{1}^{\text{You}} )</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>k_2</td>
<td>( \rho_{2}^{\text{You}} )</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>k_3</td>
<td>( \rho_{3}^{\text{You}} )</td>
</tr>
<tr>
<td>Facebook</td>
<td>B</td>
<td>k_1</td>
<td>( \rho_{1}^{\text{Ace}} )</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>k_2</td>
<td>( \rho_{2}^{\text{Ace}} )</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>k_3</td>
<td>( \rho_{3}^{\text{Ace}} )</td>
</tr>
</tbody>
</table>

![Fig. 2 An Example of FIB Table.](image)

Table 1: An Example of PHA Method

<table>
<thead>
<tr>
<th>Priority Queue</th>
<th>Players</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Q_1 )</td>
<td>( \rho_1 ) ( \rho_1^1 ) ( \rho_1^2 ) ( \rho_1^3 )</td>
</tr>
<tr>
<td>( Q_2 )</td>
<td>( \rho_2 ) ( \rho_2^1 ) ( \rho_2^2 ) ( \rho_2^3 )</td>
</tr>
<tr>
<td>( Q_3 )</td>
<td>( \rho_3 ) ( \rho_3^1 ) ( \rho_3^2 ) ( \rho_3^3 )</td>
</tr>
<tr>
<td>( Q_4 )</td>
<td>( \rho_4 ) ( \rho_4^1 ) ( \rho_4^2 ) ( \rho_4^3 )</td>
</tr>
</tbody>
</table>

4. Evaluation

In order to assess the effectiveness of our scheme for CCN, We implemented the game theoretical Interest forwarding scheme by extending ccnSim [14] simulator which is the OMNET++ based CCN simulator. We run our simulation on an Intel Core 2 Duo CPU T9400 running at 2.53 GHz and 4 GB of memory.

4.1 Simulation Settings

In simulation, a network is modeled as a graph \( G(n,p) \), where \( n \) is the number of nodes in the network and \( p \) is the probability of a connecting link exists between two nodes. GT-ITM [23] is used to generate a topology simulating the Internet, whose \( n = 50, p = 0.3 \). Links between nodes are characterized by their bandwidth and propagation delay.
The bandwidth of each link is set to 100Mbs and link propagation delays range from 1ms to 5ms.

In our network, we adopt the chunk size is 10KB; file size is about $10^3$ chunks; catalog size is up to $10^7$ files. We select cache sizes of 10 GB and keep the ratio of cache over catalog on the order of $10^{-5}$ ($\text{Cache}/\text{Catalog} = 10^{-5}$). The routers use standard replacement method LRU (evicts the least recently used packet) and decision polices ALWAYS (caches every chunk it receives) [13]. The parameters of our simulation are showed in Table 2.

<table>
<thead>
<tr>
<th>Para</th>
<th>Value</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>50</td>
<td>Number of nodes</td>
</tr>
<tr>
<td>$p$</td>
<td>0.3</td>
<td>Connectivity probability</td>
</tr>
<tr>
<td>$b$</td>
<td>100Mbs</td>
<td>Link bandwidth</td>
</tr>
<tr>
<td>$d$</td>
<td>[1,5]ms</td>
<td>Link delay</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>1</td>
<td>Content popularity distribution skewness</td>
</tr>
<tr>
<td>$q$</td>
<td>0.25</td>
<td>Content popularity distribution skewness</td>
</tr>
<tr>
<td>Chunk size</td>
<td>10KB</td>
<td>CCN chunk size</td>
</tr>
<tr>
<td>Cache size</td>
<td>10GB</td>
<td>Cache size of each node</td>
</tr>
<tr>
<td>Catalog size</td>
<td>$10^6$ files</td>
<td>each file is $10^3$ chunks</td>
</tr>
<tr>
<td>(Cache/Catalog) ratio</td>
<td>$1*10^{-5}$</td>
<td>$C/(</td>
</tr>
</tbody>
</table>

There are two repositories which store the same content. Among the nodes, we randomly select 2 nodes which are connected to repository. We use the Mandelbrot-Zipf distribution model to calculate the content popularity, where $\alpha = 1.5$ and $q=0.25$. The network has 10 client users which are connected to its border nodes. Users perform File-level requests according to a Poisson process with exponentially distributed arrival times at a 1 Hz rate.

### 4.2 Simulation Results

We do the evaluation and analyze the effectiveness of CCN with three different Interest forwarding algorithms:

- **IFD**: A node forwards the Interests by game theoretical multiple Interest Forwarding Decision method;
- **CCN-S**: A node forwards the Interests by the shortest path algorithm;
- **CCN-B**: A node forwards the Interests to all interfaces through which the Data is available.

We compare the four schemes by focusing on the metric: average data retrieve time, which denotes the user’s benefits directly.

Fig. 3 shows data retrieve time as function of cache over catalog ratio with content popularity distribution skewness $\alpha = 0.8$ in CCN with three different Interest forwarding methods. Abscissa is the cache over catalog ratio. Ordinate is the average data retrieve time. We can see that with the cache size increases, data retrieve time sharply decreases. When the cache size is small, the IFD has slightly better performance than CCN-S. However, as the cache size increases, the gap between three forwarding mechanisms is becoming smaller until same. IFD has dramatically better performance than original CCN-B. This is due to the fact that CCN forwards Interest to all reachable service instances, which makes up the large of bandwidth and makes the network congestion.

![Fig. 3 Data retrieve time as function of cache over catalog ratio.](image1)

![Fig. 4 Data retrieve time as function of content popularity skewness [\(\alpha\)].](image2)
Fig. 4 depicts the Data retrieve time as function of content popularity skewness $\alpha$ with cache size $C = 10\text{GB}$. It can be seen that data retrieve time decreases as the content popularity distribution skewness alpha increases, especially when alpha more than 1.0, there is a sharply decline. CCN with IFD has similar performance with CCN-S when the skewness $\alpha$ is small. As skewness $\alpha$ increase, IFD has better performance than CCN-S. This is because that IFD forwards the Interest to multiple paths which can get higher cache hits than CCN-S when the popular data increase.

Fig. 5 Cache hit ratio as function of cache over catalog ratio

We also evaluate the cache hit cache hit ratio as function of cache over catalog ratio for three forwarding schemes. As showed in Fig. 5, with the increase of cache over catalog ratio, the cache hit radio of all schemes increased. Furthermore, IFD scheme has higher cache hit ratio than the other two schemes when cache over catalog ratio is smaller than $10^{-3}$; when cache size over catalog ratio is bigger than $10^{-3}$, IFD scheme has lower cache hit than CCN-B, but better performance than CCN-S.

5. Conclusions

This paper investigates Interest forwarding strategy in Content-Centric Networking where a set of Interests sharing a multiple interfaces from which the Interest can get the response from repository. Users are assumed to be self-regarding and make their decisions with the sole goal of maximizing their perceived quality. We presented a game theoretical multiple Interest Forwarding Decision (IFD) method to improve the users’ payoff when the network is not in the high traffic. IFD used non-cooperative game theory to analysis the multiple Interests forwarding decision. We took the Interest flow $I$ as the game player. Each game player maximizes his payoff cost. In the network perspective, we proposed a Potential Heuristic Allocation (PHA) method to queue the replica Interests which considers the fairness and network efficiency simultaneously. IFD improved the utilization rate of network resources.

We did evaluation for CCN with three different Interest forwarding methods. The simulation results show that our proposals improved the CCN performance. It can be adaptively make the multiple Interest forwarding decisions in different network traffic scenarios.

In the future, we are planning to discuss different game theory models for Interest forwarding decisions in CCN. Furthermore, we will consider the multipath Interest forwarding for CCN.

References


Chengming Li received the B.S. degree in software engineering from Dalian University of Technology and M.S. degree in computer application technology from Dalian University of Technology in 2009 and 2011, respectively. Now he is a Ph.D. candidate in Kyushu University, supported by China Governmental Scholarship. His research interests include virtualization technologies and future internet.

Koji Okamura is a Professor at Department of Advanced Information Technology and also at Computer Center, Kyushu University, Japan. He received B.S. and M.S. Degree in Computer Science and Communication Engineering and Ph.D. in Graduate School of Information Science and Electrical Engineering from Kyushu University, Japan in 1988, 1990 and 1998, respectively. He has been a researcher of MITSUBISHI Electronics Corporation, Japan for several years and has been a Research Associate at the Graduate School of Information Science, Nara Institute of Science and Technology, Japan and Computer Center, Kobe University, Japan. He is interested in Internet and Next Generation Internet, Multimedia Communication and Processing, Multicast/IPv6/QoS, Human Communications over Internet and Active Networks. He is a member of WIDE, ITRC, GENKAI, HIIK projects and Key person of Core University Program on Next Generation Internet between Japan and Korea sponsored by JSPS/KOSEF.
Modeling function calls in program control flow in terms of Petri Nets

Dmitriy Kharitonov$^1$ and George Tarasov$^{1,2}$

$^1$ Institute of Automation and Control Processes
Far-Eastern Branch of Russian Academy of Sciences
Vladivostok, 5 Radio st, 690041, Russia
demiurg@dvo.ru

$^2$ Far-Eastern Federal University
Vladivostok, 8 Suhanova st, 690950, Russia
george@dvo.ru

Abstract
This article presents a method for representing the C/C++ function call in terms of compositional Petri Nets. Principles of modeling function and function call in the program are described. Formal composition operations to construct program model from models of its functions and modules are also introduced. All results are illustrated on an example of a real parallel program.

Keywords: Compositional Petri Net, Control Flow, Program Model, C/C++ Programming Language.

1. Introduction

There are many subjects in computer science, such as static and dynamic analysis [1], program verification [2], performance analysis and debugging [3], which are based on the software program models. Construction of the model is very difficult and tedious process. It should represent all aspects of the source program that satisfy the goals of modeling, and at the same time, it should not be too much detailed for analysis methods to be applied with useful results [4]. One should mention that model checking technique has the state space explosion problem [5]. This problem makes impossible full verification of the program, but part-by-part program verification can be done using special state space reduction methods or reducing model itself, removing unimportant details. The interested area of research is to develop methods and algorithms for automatic and semi-automatic proper model synthesis [6].

In this paper authors consider a technique to represent function calls in the control flow model of imperative programs. Representation includes actions that are performed on a both caller and callee sides. Proposed technique has two main features. At first it implies an easy possibility to develop an automatic model generator from real program source code. And in second it eliminates redundant states for reusable code parts in resulting model. Function models are aggregated in modules that can call other module functions. The whole program model is constructed as composition of the set of modules and the program entry point.

Petri nets are quite often used for modeling control flow in imperative programs [7, 8, 9, and 10]. Each transition is the action in program code that may contain a number of some operators; commands or blocks (depend on details). Each place is the state of the computational process between two actions. Often state means a current program memory values. Tokens are characterized the state of computational process execution. Their location in the places determine the execution of various actions, hence the place is often considered as a pre- and post-conditions of actions. Rules of transition firing let to visualize naturally the control flow of basic algorithmic constructions of imperative programs: condition, cycle and switch. We use concept Petri net object (or PN-object), introduced in [11] to construct program model. Each function of each module and the entire program as a whole represent an individual PN-object. Call from one PN-object to another describes control flow transfer from caller function to callee one.

Further, in this paper we give a brief background theory of PN-object calculus. After that the principles of modeling function and function call in the program are described. Then the technique of constructing a model of the program module and a complete model of program is presented. Finally, an example of a parallel program and its model in terms of PN-object is given.

2. Petri net object

Let $A = \{a_1, a_2, \ldots, a_k\}$ is a set. Multiset on $A$ is defined as a function $\mu: A \rightarrow \{0, 1, 2, \ldots\}$, that associates with each element of set $A$ some non-negative integer number. Multisets are conveniently written as a formal sum...
The name of the access point is represented as a number of tokens inside places. Object’s output incidence functions.

The following graphical notation is used for objects. PN-object is an initial marking. Also we denote if there are no transitions for each . We say , if , and . If , then this multiset is denoted as . Also we denote if . Set of all finite multisets on set is denoted as .

Definition 1: Petri net is a tuple , where
1. is a finite set of places;
2. is a finite set of transitions, with ;
3. : is an incoming incidence function; and
4. : is an outgoing incidence function.

Multisets are referred to as incoming and outgoing multisets of places for transition .

We will use standard graphical notation of Petri nets as a bipartite directed graph, where the places are represented by circles, and transitions – by rectangles. Places and transitions are connected by arcs representing the input and output incidence functions.

Definition 2: Petri net object (here and after PN-object or just object for short) is a tuple , where
1. is a alphabet and ;
2. is a set of access points (AP), each having form ;
   • is a name of access point,
   • is a alphabet and
   • is a transition labeling function;
3. is an initial marking.

Less formally PN-object is a Petri net, provided with a set of labeling functions.

To designate objects and their access points we will use capital letters and Greek letters . Entry means object with two access points and . The following graphical notation is used for objects. PN-object is drawn as rectangle, and its structure, if necessary and possible, is displayed inside a rectangle. The marking is represented as a number of tokens inside places. Object’s access points are displayed as small squares on the boundary of the rectangle. The name of the access point is placed near the square (if needed alphabet is given). The label of transition is shown inside or near the transition and consists of access point name followed by colon and then label from alphabet. If it is clear from context, the access point name is not specified.

To refer corresponding components of the access point next designation will be used: , . The label of transition in access point will be defined as .

id is the name that formed from simple alphabet. Examples of names: , , , and other.

Note 1: Further in this article we will consider only those objects that have transitions labeled by the only one labeling function:

Let’s define a number of operations that will be used later in the functions models construction.

Definition 3: Consider two objects , and , where . Form of union of objects creates new object where .

Definition 4: Consider PN-object where have two access points and , and . Operation of union of access points and creates new PN-object where .

Some simple algorithms are acceptable for further operations:

Operation of union of access points instead of two access points creates a new one that combines their alphabet and labeling function.

Definition 5: Consider PN-object where . Operation of restriction of object by access point creates new object where .

Copyright (c) 2014 Advances in Computer Science: an International Journal. All Rights Reserved.
Restiction of object by access point deletes each transition having label from $\Delta(a)$ with all adjacent arcs.

**Definition 6:** Consider PN-object $E_1 = (\Sigma_1, \Gamma_1, M_{o1})$ and its access points $\alpha, \beta \in \Gamma_1$, where $\alpha = (id, a, \Delta_1, \sigma_1)$, $\beta = (id, b, \Delta_0, \sigma_0)$ and $\Sigma_1 = (S_1, T_1, *_{\{1\}}).$ Operation of simple composition of object $E_1$ by access points $\alpha$ and $\beta$ forms new object $E = (\Sigma, \Gamma, M_0)$, where $\Gamma = \Gamma_1, M_0 = M_{o1}$ and $\Sigma = (S, T, *_{\{1\}}, O_{*}):$

1. $S = S_1.$
2. $T = T_1 \cup T_{syn}$ where $T_{syn} = \{ \mu_1 + \mu_2 | \mu_1, \mu_2 \in M(T_1) \}$ and $\sigma_\alpha(\mu_1) = \sigma_\beta(\mu_2) > 0$, sum $\mu_1 + \mu_2$ is minimal, i.e. no sum exists $\mu'_1 + \mu'_2$, that $\mu'_1 + \mu'_2 < \mu_1 + \mu_2$ and $\sigma_\alpha(\mu'_1) = \sigma_\beta(\mu'_2)$.
3. $O_{*_{\{1\}}} = \{ O_{_{\{1\}}} \cup \{ \mu_1 + \mu_2, *_{\{1\}}(\mu_1) + *_{\{1\}}(\mu_2) \} | \mu_1 + \mu_2 \in T_1, \mu_1, \mu_2 \in M(T_1) \}$.
4. $O_{*_{\{2\}}} = \{ O_{_{\{2\}}} \cup \{ \mu_1 + \mu_2, (\mu_1)^+ \ast (*_{\{2\}}, \mu_2)^+ \ast \} | \mu_1 + \mu_2 \in T, \mu_1, \mu_2 \in M(T_1) \}$.
5. $\forall t \in T_{syn} \forall \xi \in \Gamma : \xi(t) = 0.$

Operation of simple composition for one object (unary form) and for two objects (binary form) is denoted accordingly

$$E = [E_1]^\beta. \quad E = E_1 a|_\beta E_2 \equiv [E_1 \oplus E_2]^\beta.$$  

Operation of simple composition adds to object $E_1$ a number of new synchronization transitions $T_{syn}.$ New transition are defined by multisets of symbols $\mu_1 + \mu_2$, where $\mu_1, \mu_2 \in M(T)$ and have no labels. Incoming and outgoing multisets of the new transitions are calculated accordingly:

$$*_{\{1\}}(\mu_1 + \mu_2) = *_{\{1\}}(\mu_1) + *_{\{1\}}(\mu_2). \quad (\mu_1 + \mu_2)^* = (\mu_1)^* + (\mu_2)^*.$$  

**Definition 7:** Consider PN-object $E_1 = (\Sigma_1, \Gamma_1, M_{o1})$ and its access points $\alpha, \beta \in \Gamma_1$. Operation of directional composition (unary form) of object $E_1$ by access points $\alpha$ and $\beta$ creates a new object $E$, that

$$E = a_{\{1\}}^\beta = \partial_\alpha([E_1]^\beta).$$

For two objects $E_1 = (\Sigma_1, \Gamma_1, M_{o1})$ and $E_2 = (\Sigma_2, \Gamma_2, M_{o2})$ with access points $\alpha \in \Gamma_1$ and $\beta \in \Gamma_2$ accordingly, operation of directional composition (binary form) is denoted as

$$E = E_1 \rightarrow E_2 = \partial_\alpha(E_1 a|_\beta E_2).$$

This definition implies that a unary directional composition performs two operations on an object simple composition by $\partial_\alpha$ and $\beta$ and restriction by $\partial_\alpha$. Binary directional composition expressed in terms of unary, having done a formal union of the two original objects.

Let us formulate some properties that operations on the PN-objects have. We emphasize that properties are true in the context of the Note 1. For the cases of arbitrary transitions labeling, these properties may not hold.

**Statement 1:** Some properties of the operations on a PN-objects.

1. Operation of formal object union is commutative and associative:

   $$E_1 \oplus E_2 = E_2 \oplus E_1,$$
   $$E_1 \oplus (E_2 \oplus E_3) = (E_1 \oplus E_2) \oplus E_3.$$  

   **Proof:** Follows from the commutativity and associativity of union of sets and multisets operations and definition 3.

2. Union of access points operation is commutative and associative:

   $$(E_1)_{\gamma_{\alpha \beta}} = (E_2)_{\gamma_{\beta \alpha}},$$
   $$(E_1)(E_2)_{\gamma_{\alpha \beta}} = (E_1)_{\gamma_{\beta}}(E_2)_{\gamma_{\alpha}}.$$  

   **Proof:** Follows from the commutativity and associativity of union of sets and multisets operations and definition 4.

3. Associativity of the restriction by access point operation:

   $$\partial_\alpha(\partial_\beta(E)) = \partial_\beta(\partial_\alpha(E)).$$

   **Proof:** Follows from definition 5. As far as subsets of labeled transitions after restrictions by $\alpha$ and $\beta$ do not intersect, then in both cases we get the same set $T$, and hence the equality is fair.

4. For two objects $E_1 = (\Sigma_1, \Gamma_1, M_{o1})$ and $E_2 = (\Sigma_2, \Gamma_2, M_{o2})$ and their access points $\gamma \in \Gamma_1$ and $\alpha, \xi \in \Gamma_2$ next equality holds:

   $$E = E_1 \{ \gamma \} \partial_\gamma(E_2) = \partial_\gamma(E_1 \{ \gamma \} E_2).$$

   **Proof:** Follows from definitions 3 and 5.

5. Associativity of directional objects composition:

   $$\alpha \beta \beta \alpha \gamma \gamma \gamma \gamma \gamma$$

   **Proof:** Follows from definitions 3 and 5.

   $$E_1 \rightarrow E_2 \rightarrow E_3 = E_2 \rightarrow E_3 \rightarrow E_1.$$  

   $\gamma \gamma \gamma \gamma \gamma$
Associativity of union of access points and directional composition of objects operations make it possible not to take into account the order of these operations on the set of original objects. As a result, to denote operation of directional composition of the object \( \alpha \), where the set of access points consists from \( \Pi = \{ \pi_1, ..., \pi_n \} \) and the access point \( \alpha \), the next notation will be used:

\[
E_{\alpha}(\alpha) = (\pi_1, ..., \pi_n) \rightarrow (\pi_1) = \alpha.
\]

We will use the following graphical notation to display PN-object operations. In a formal union of objects inside the result PN-object the original PN-objects are placed. All access points of internal objects are duplicated on the border of the external object connected with the appropriate lines. Access points for restriction operation are drawn fully filled. Operation of directional composition is represented by arrow from one access point to another. The source access point by which restriction operation is performed is fully filled. Arrow indicates the direction of synchronization: events in object \( E_1 \) can occur only if there are equivalent events in object \( E_2 \).

On the Figure 1 a sequence of operations compounded operation of directional composition of two objects \( E_1 \) and \( E_2 \) is shown. This sequence consists of formal union of source objects, simple composition by \( \alpha \) and \( \beta \) access points, and restriction by \( \alpha \) access point. The original compositional representation and result of each sub-operation are shown on a Fig. 1a)–1d) respectively.

4. Model of function and function call in imperative program

In this section we will use mathematical operations on objects described above to construct the model of the function in imperative programming languages. In such languages the concept of a function is often associated with traditional structured programming concept subroutine. Function is a certain sequence of main program actions segregated to perform the repetitive calculations. In modern programming languages the terminology associated with the concept of a function has become quite blurred. Different programming languages use different synonyms to identify the same entity: function, procedure, method, subroutine, subprogram, etc. Despite the number of names and according syntactic and semantic differences, the essence of the function (as a subroutine) remains the same and is a sequential execution of the next set of steps:

1. Special command function call transfers the control flow to the function while execution of commands following the call is temporarily suspended;
2. Command that forms function body run until come across special command return from function;
3. Function execution completed and control flow transfers to the next command that follows the function call.

Let’s give a formal definition of the basic concepts models: function call and function.

**Definition 8**: Let us given a PN-object \( E = (\Sigma, \Gamma, M_0) \), \( \Sigma = (S, T, \ast, O, \ast\ast) \) and alphabet \( \Delta_f = \{ \text{begin}_f, \text{end}_f \} \). Object \( E \) has access point \( \alpha \in \Gamma \), in form \( \alpha = (f, \Delta_f, \sigma_f) \).

Then three elements \( (t_1, s, t_2) \) in object \( E \) structure consisting of two transitions and places \( t_1, t_2 \in T, s \in S \) that

1. \( t_1 \neq t_2 \),
2. \( t_1 \ast = s = \ast t_2, s\ast = t_1, \ast s = t_2 \),

will be called **model of the call of function \( f \) in the context of access point \( \alpha \), if**

\[
\sigma_f(t_1) = \text{begin}_f, \sigma_f(t_2) = \text{end}_f.
\]

Less formally model of function call will denote three marked and connected in certain way net elements, where:

- first transition models control flow transfer to the function (label \( \text{begin}_f \)), this transition is called **call transition**;
- place models a state of function completion waiting;
- second transition models control flow return from the function (label \( \text{end}_f \)), this transition is called **return transition**.

This definition naturally describes the calls of the same function several times and the calls of several different functions in a single PN-object. In the first case, each call in the structure of the object \( E \) will have its own three elements with identical labels. In the second case, each unique function \( f \) call is described by a separate alphabet \( \Delta_f \) and a separate access point \( \sigma_f \in \Gamma \). There is general case possible, where the object \( E \) has one access point describing the calls of all functions. Using operation of union of the access points we can join all access points to one access point named, for example, **calls**. This case is shown on a Fig. 2.

Let’s define the function model, assuming that the internal structure of the function control flow is already described by some Petri net.

**Definition 9**: Let us given PN-object \( E_f = (\Sigma_f, \Gamma_f, M_{0f}) \), where \( \Sigma_f = (S_f, T_f, \ast, \Omega_f, \ast\ast) \) – object structure, describing control flow of the function body, and there is in \( \Gamma_f \) two disjoint subsets of access points

\[
(\text{IN}_f \cup \text{OUT}_f) \subseteq \Gamma_f, \quad \text{IN}_f \cap \text{OUT}_f = \emptyset.
\]

\( \text{IN}_f = \{ \text{in} \} \) is a subset of incoming access points, consisting of one access point \( \text{in} = (f, \Delta_f, \sigma_f) \). \( \Delta_f = \{ \text{begin}_f, \text{end}_f \} \). \( \text{OUT}_f = \{ \text{out}_1, ..., \text{out}_n \} \) is a subset of outgoing access points, satisfying the definition 8. If for the object structure \( \Sigma_f \) and for access points \( \text{in}, \text{out}_1, ..., \text{out}_n \) next statements are fair:

1. \( \exists! t_b \in T : t_b = \emptyset \land \ln(t_b) = \text{begin}_f \);
2. \( \forall t \in T_c : t \neq t_b \land t\ast = \emptyset \land \ln(t) = \text{end}_f \), where \( T_c \subseteq T \) and \( T_c \neq \emptyset \);
3. \( \forall \text{out}_i \in \text{OUT}_f : \text{out}_i(t_b) = \text{begin}_f \Rightarrow \ast(t_b) = t_1 \Rightarrow \text{out}_i(t_1) = \text{end}_f \);

then this PN-object \( E_f \) will be referred as **function \( f \) model**.

Let’s give a PN-object \( E_f \) in form \( \alpha = (f, \Delta_f, \sigma_f) \) and for access points \( \{ \text{in}, \text{out}_1, ..., \text{out}_n \} \) and for one access point named, for example, \( \text{calls} \) structure

\[
\gamma = \alpha + \beta \quad \text{id}(\gamma) = \text{calls}
\]

2. Examples of function call models.

Statements 1–3 in definition 9 mean that:

1. There is only one transition \( t_b \), that have no incoming arcs, this transition is called **incoming transition**;
2. Object structure have non-empty set of transitions \( T_c \), having no output arcs, this transitions are called **outgoing transitions**;
3. All outgoing transitions are describing calls of functions another than \( f \).

Let’s use additional notations:

\[
\text{In}(E_f) = \text{IN}_f - \text{function In returns a set of incoming access points of PN-object, Out}(E_f) = \text{OUT}_f - \text{function Out returns, accordingly, a set of outgoing access points of PN-object.}
\]

Examples of function models and function call models are shown on a Fig. 3. Objects \( E_f, E_{f1} \) and \( E_{f2} \) represent models of three different functions \( f, f_1 \) and \( f_2 \) accordingly. Each object has one incoming access point \( \text{in}_f \), from which control flow is transfered to function body, modeled by object. Object \( E_f \) has two outgoing access points describing calls of functions \( f_1 \) and \( f_2 \). Link between caller and called functions is defined through directional composition operation by proper access points.
Process of control flow transfer can be described in a next way. Consider object \( E_f \) in the state \( M = (0,1,0,0,0,0,0) \) (token in place \( s_2 \)). Then, in object \( E_f \) there can be fired transition \( t_2 \), labeled \( \text{begin}_{f_1} \). In accordance with the rules of directional composition operation, objects \( E_f \) and \( E_{f_1} \) are combined together, and from transitions \( t_3 (E_f) \) and \( t_{b2} (E_{f_1}) \) a new transition is formed, that removes token from \( s_2 \) and puts one token to the places \( s_2 (E_f) \) and \( s_9 (E_{f_1}) \). Token in place \( s_2 \) means waiting of control flow return from function \( E_{f_1} \) (awaiting token), and token in place \( s_9 \) initiates execution of function body. Awaiting token destructs, when return transition \( t_4 \) fires, that likewise connected to outcoming transition \( t_{e4} \) in object \( E_{f_1} \).

**Definition 10:** Module is a PN-object \( E = (\Sigma, \Gamma, M_0) \), that have defined functions \( In \) and \( Out \). Without loss of generality, it is assumed that
\[
In(E) \cup Out(E) = \Gamma, \vert In(E) \vert = 1, \vert Out(E) \vert \geq 0.
\]

Minimal module is considered to consist of only one function.

Less formally module is an object, that have only one incoming access point, via which can be invoked components of the module, and may have zero or more outcoming access points, via which it connects to other modules. Module, having empty set of outcoming access points is referred to as full module.

Let’s introduce the operation of modules composition as a method for designing complex modules from a set of more simple ones. Let’s denote for a module \( F \) subset \( \text{SELF} \subset \Gamma \) that includes all output access points of the module, having alphabets included in the alphabet of input access point for this module:
\[
\text{SELF} = \{\text{out}_i \mid \text{out}_i \in Out(F) \land \Delta(\text{out}_i) \subseteq \Delta(In(F))\}.
\]

Access point belonging to the set \( \text{SELF} \) will be called the internal access points.

**Definition 11:** Let’s \( E_1 \) and \( E_2 \) are modules. Then module composition operation builds from two modules \( E_1 \) and \( E_2 \) the new module \( E = (\Sigma, \Gamma, M, 0) \) such that:
\[
E = E_1 \cup E_2 = (E_1 \cup E_2)[\text{in}\rightarrow \text{in}^*],
\]
where
\[
\text{In}(E) = \{\text{in}\} \text{ incoming access point of new module},
\]
\[
\text{Out}(E) = (\text{Out}(E_1) \cup \text{Out}(E_2)) \setminus \text{SELF} - \{\text{set of new module outcoming access points},
\]
\[
in_{i1} \text{ and } in_{i2} - \text{ incoming access points of PN-objects } E_1 \text{ and } E_2 \text{ accordingly.}
\]

Let’s consider an example of two modules composition (fig. 4) \( M_1 \) and \( M_2 \), that has input access point \( in_{i1} \) and \( in_{i2} \) and subsets of outcoming access points \( \{f_1, f_3\} \) and \( \{f_1, f_2, g\} \) accordingly. It is assumed that via access point \( g \) a function in the second module is called from the first module. Figure 4(a) shows source modules and their access points. Figure 4(b) shows composition of source modules, specifically all the transformations of the source objects: objects are formally united in one object \( M \), and union of access points \( in_{i1} \) and \( in_{i2} \) is performed, subset of internal access points \( \text{SELF} \) is highlighted, consisting of \( \{g\} \), than directional composition by access points from sets \( \text{SELF} = \{g\} \) and \( \text{In} = \{\text{in}\} \) is executed. As the result (fig. 4(c)) a new module \( M \) is obtained, that have all access point of source objects except for \( g \).
Thus, applying the operation of modules composition to a certain set of program functions models one module can be obtained that contains a single incoming access point to call any function from the set and a set of output access points to call external functions. The presence or absence of the outcoming access points in the module obtained depends upon our considerations of nested function calls. For example figure 4 shows source modules having calls to some functions ($f_1$, $f_2$, and $f_3$). And there can be two analysis cases. In one case functions $f_1$, $f_2$ and $f_3$ are considered in the same way as all others and we need to include their models into the whole program model. But another way we can consider these functions as elementary, not worth scrutiny from the point of program control flow modeling. In that case each of these functions can be modeled by one transition without awaiting places, access points and so on. And thus the result of composition will be full module having only one incoming AP. The level of details is chosen by man and depends on modeling purposes.

Let’s consider final program model, which execution in imperative languages begins from some start function (program entrance point). Before the control flow is given to this function, there is a special imperative block completed from executable file - loader, that compiler adds to the executable image automatically. The role of the loader is to initialize the system environment of the program and then calling the starting function.

**Definition 12:** Let us given PN-object $E = \langle \Sigma, \Gamma, M_0 \rangle$, where $\Sigma = (S, T, \ast, \ast, \ast, \ast)$ and $\Gamma = \{\alpha\}$, $\ast = (\text{out}, \Delta, \sigma)$ and $\Delta= \{\text{begin, end}\}$. We will call $E$ the model of loader, if:

- $\exists s_b, s_e \in S : s_b \neq s_e$,
- $s_b = \emptyset$, $s_e^* = \emptyset$,
- $\forall s \in S : M_0(s) = \begin{cases} 1, & \text{if } s = s_b \\ 0, & \text{otherwise} \end{cases}$,
- $\exists(t_1, s, t_2), t_1, t_2 \in T, s \in S : s \neq s_b \neq s_e \land t_1^* = s = t_2^*$ and $\text{out}(t_1) = \text{begin}$ and $\text{out}(t_2) = \text{end}$.

Less formally the loader can be defined as some PN-object with highlighted initial $s_b$ and final $s_e$ places, modeling, respectively, the beginning and the end of the program, with a single start function call via the access point $\text{out}$ and initial marking, having token only in the initial place $s_b$.

**Definition 13:** Let us given loader $E_1$ and full module $E_2$ with access points out and in such that:
- $\text{Out}(E_1) = \{\alpha\}$, $\text{In}(E) = \{\beta\}$, $\alpha = (\text{out}, \Delta, \sigma, \beta), \beta = (\text{in}, \Delta_\beta, \sigma_\beta)$.
- $\Delta_\alpha = \Delta_\beta = \{\text{begin, end}\}$.

Then object $E = \partial_\beta(E_1 \rightarrow E_2)$ will be called model of imperative program.

Example of imperative program model in terms of PN-objects is shown on a Fig. 5(b). Directional object composition operation joins these objects using access points. Restriction operation removes remaining access point in from model of module from it useless. The result is an object with no access points, the structure of which models the control flow of the program.
5. An example of imperative C++ program model

Let’s use depicted above mathematical apparatus for constructing a model of the program, written in an imperative programming language. As an example, we consider the C language, as one of the most popular general purpose programming languages, designed for a wide range of applications – from performing simple data processing and up to creation of operating systems. Briefly it can be characterized as an imperative, structured programming language. One of its distinctive features is the presence of the only startup function `main`, which is the entry point of the program, which begins execution of program instructions. In addition to the function `main`, a C program can contain any number (within an acceptable by hardware and operating system range) of other functions, with mutually different names. Nesting function descriptions are not allowed, i.e., all functions are equivalent and they can be accessed from anywhere in the program.

Let’s briefly describe the scope of concepts associated with the function in C. Description of each function consists of the function header and its body. The function header describes such features as the function name, return type, and a list of the input (and output) parameters. Description of the function body follows next to the description of the function header. Body of the function is a block of statements that consist of next basic algorithmic constructions: expressions, conditions, cycles, switches, the function call operators, return from function operators, and others. Execution of statements is performed sequentially until the last statement executed, or the return from function operator meet. The transfer of the control flow to the body of the function is carried out by the construction `function call`, which is partially supported on a hardware level of CPU.

The figure 6 shows an example of the “set dividing program” and a model of its control flow. Model consists of five PN-objects. Object $E_{\text{loader}}$ is the program loader. Initial state of the program is shown by two tokens. Each token is associated with a process, which by program design either collects larger elements of two sets or smaller ones. Loader executes just one command - call the `main` function via transitions, labeled `begin_{\text{main}}` and `end_{\text{main}}` accordingly. $E_{\text{main}}$ object simulates common actions of each process. Transition labeled `INIT` matches strings 33–35 of the program. After initial initialization a choice construction is performed that correspond to the rows 36–37 of the program. That construction models all possible cases of further processes execution, more precisely, either control flow transfer to object $E_{\text{small}}$ (function `Small` call), or control flow transfer to object $E_{\text{large}}$ (function `Large` call), or execution of bogus action `GOTO`, which is possible, if none of the conditions are satisfied.

In terms of the used Petri nets notation any of the tokens may initiate any of the transitions labeled `begin_{\text{small}}`, `begin_{\text{large}}` or `GOTO`. However, we assume that this choice can be determined, and one piece will excite a transition `begin_{\text{small}}`, while another - `begin_{\text{large}}` . Accordingly, function calls constructions will give control to the objects $E_{\text{small}}$ and $E_{\text{large}}$. These objects have similar structures of the control flows and perform the same type of actions that differ by labels. Lines 16–23 of the program match the object $E_{\text{small}}$ structure, lines 24–31 - the object $E_{\text{large}}$ structure. Both objects have several calls to common interaction function `SendRecv`, the control flow of which is modeled by object $E_{\text{SR}}$. Rules for firing the transition `SENDRECV` in this example are not regulated. It is
assumed that both tokens (both processes) excite it in accordance with the general rules of transitions firing. Synchronization of the processes can be described using function call construction and the model of the function MPI_SendRecv as one more PN-object.

6. Conclusions

This paper presents a method for constructing a model of imperative program control flow as the composition of control flow models of its component functions.

Using the notation of PN-object function call and function body is described. Formal directional composition of objects operation defined, that allows to obtain single PN-object modeling control flow of program from a number of smaller PN-object. Representation of the control flow in the form of function models composition have a number of significant advantages with respect to the flatten representation of one model.

In particular it partially simplifies the states explosion problem, because the number of states in the model is reducing a lot due to the fact that there are no duplicated states of the same function called from different places of the program. Stack of function calls in each moment of time is reflected by a waiting places in function call constructions. Then the procedure of automatic model generation from source code is simplified, because each function in program maps to its own PN-object, connected to others by a set of links. And the last advantage is that modularity of model gives more flexibility in model analysis. Input of analysis algorithms may have full compositional representation, flatten representation of program or a model of subset of program function realizing logical part or functionality of the whole program.

The results obtained give us an opportunity for further successful development of the submitted approach. In particular, the work does not address how to describe recursive functions, although quick examination of the PN-objects definition and operations on them are quite suitable for compositional representation of recursion in imperative programs. Virtual functions in object-oriented programming languages are also not considered, and relate to runtime type recognition. Also there is a need to study individually the questions of description data and methods of data transfers back and forth to function.

Acknowledgments

This work was supported by the Russian Academy of Sciences (program Num. 43 of the Presidium of the RAS “Fundamental problems of the mathematical modeling”), by the Far Eastern Branch of the Russian Academy of Sciences, and by the Far Eastern Federal University.

References


Fig. 6. Example of program and its model in terms of Petri Nets.


Dmitry Kharitonov is a senior research fellow at the Institute of Automation and Control Processes of the Far-Eastern Branch of Russian Academy of Sciences. He holds his PhD in Software of Computational Networks and Systems (Candidate of Technical Sciences, Vladivostok, Russia, 2001). He holds his MSc in Applied Mathematics (Moscow Institute of Physics and Technologies). He is interested in program verification, program modeling, parallel programming languages and translators.

George Tarasov is a research fellow at the Institute of Automation and Control Processes of the Far-Eastern Branch of Russian Academy of Sciences and Far-Eastern Federal University. He holds his MSc in Automation Control Systems (Far-Eastern Federal University, Vladivostok, Russia, 1999). He is interested in parallel programming, program verification and performance analysis.
On the Layer based Seamless Handover Schemes for Mobile Data Network

Y. J. Lee
Department of Technology Education, Korea National University of Education
Cheongju, 363-791, South Korea
lyj@knue.ac.kr

Abstract
This paper presents seamless handover schemes survey based on layer approach. Efficient handover scheme is very important in order to minimize the handover latency in mobile network. Latency is mainly affected when new IP address has to be obtained in layer 3 and layer 4 handover mechanisms because this is processed by software. Therefore, we have to focus on the exact mechanism to obtain new IP address and delete old IP address. IPv6 tries to reduce the latency by using stateless-auto-configuration method instead of DHCP (Dynamic Host Configuration Protocol) in layer 3 handover. However, this method limits most of applications using IPv4. SCTP (Stream Control Transmission Protocol) uses multi-homing features with dual interfaces. L4 handover scheme using SCTP is well known to reduce the handover latency more than the handover scheme using TCP. However, SCTP handover scheme utilizes IP address in L4, which violates the strict layer independence. This paper provides us with the pros and cons to determine which handover scheme is applicable to the specific application environment.

Keywords: Handover schemes, Layer based approach, SCTP, Data Network.

1. Introduction
Handover scheme is one of very important issues in data communication networks. These schemes are classified into four categories: Layer 1(L1), Layer 2(L2), Layer 3(L3), and Layer 4 (L4) handover.

L1 handover performed in the physical layer mainly focuses on the change of signal power. That is, the main problem is how to detect the real signal power in order to handover to new AP (Access Point).

L2 handover performed in data link layer compose of three procedures. Movement detection is to first check the reachability of old AP and then discover new AP based on the beacon signal power and SIR (Signal-to-Interference Ratio). If necessary, L2 handover selects new AP and makes a reassociation for new AP through authentication.

L3 handover performed in the network layer conducts another movement detection: First, L3 handover investigates the reachability of old AR (Access Router), then checks the validity of current AR, and then discovers new AR. L3 handover then selects new AR (if necessary). Finally, it conducts DAD (Dynamic Address Determination) [1]. L3 handover configures new IP (Internet Protocol) address based on stateless auto-configuration with IPv6 or DHCP (Dynamic Host Configuration Protocol) [2] with IPv4/IPv6 and NAT (Network Address Translator) with IPv4. L3 handover also makes a registration with new AR using BU (binding update).

L4 handover performed in the transport layer conducts the following five steps: MN (Mobile Node) adds new IP to IP_list. MN sends ASCONF_add chunk to CN (Correspondent Node) and receives ASCONF_ACK chunk. And it then sends ASCONF_set_primary chunk to CN. Finally, MN updates IP_list [3].

Now, we investigate the handover schemes in detail based on the movement detection, the configuration IP address, and the binding updates. For the movement detection, IPv4 first check the reachability of old AR by using ARP (Address Request Protocol) request/reply mechanism [4]. It then checks the validity of old AR by using unicast RS (Route Solicitation)/RA (Route Advertisement) mechanism. And it then tries to discover new AR by using all-router multicast RS/RA [5] and then selects new AR. It finally performs DAD by sending the ARP multicast for all subnet. This procedure is the same in both IPv4 and IPv6.

In order to configure new IP address, IPv4 uses the DHCPv4 and link-local-IP (NAT). On the other hand, IPv6 uses DHCPv6 and stateless-auto-configuration-link-local-site-local-global mechanism [6]. For the binding update, MIPv4 (Mobile IPv4) uses HA (Home Agent) and FA (Foreign Agent) and CN. MIPv4 can also use the
route optimization option. MIPv6 uses the route optimization as a default.

Seamless IP mobility is composed of two categories: One is the local mobility management mechanism and another is fast handover mechanism. Local mobility management scheme is classified into hierarchical mobile IP and independent region movement. In the hierarchical mobile IP scheme, MIPv4 uses regional registration mechanism. Meanwhile, MIPv6 uses HMIPv6 (Hierarchical Mobile IPv6 Mobility Management) mechanism. In the independent region movement scheme, MIPv4 uses mobile IP for global mobility, and host based routing mechanism such as Celluar and Hawaii [7]. Fast handover uses L2 trigger. MIPv4 uses the low latency handoffs and MIPv6 uses the anticipated handover or tunnel based handover mechanism.

The focus of this paper is to investigate several handover schemes to minimize the latency during handover process. Furthermore, this paper studies the factors to affect the seamless mobility.

We describe the Layer 2, 3, and 4 handover schemes in Sections 2, 3 and 4 respectively. We conclude this paper in Section 5.

2. Layer 2 handover schemes

We first describe the signal power detection problem and then investigate the L2 handover mechanism.

The important problem of L2 handover is how to detect the real signal power to indicate the real handover. The use of the signal strength as handover trigger requires the definition of a threshold. However, the signal strength threshold is a system-specific parameter. Moreover, the threshold might be even specific for devices from different vendors.

The signal strength usually includes interference. Consequently, in environments with high interference, the channel might be bad although the measured signal strength indicates a good channel. For example in IEEE 802.11 wireless LANs, the access point transmits a link layer beacon that can be used for signal strength measurements. The received signal strength may change rapidly. A typical, well-known example for such a scenario is the corner effect. In order to include the signals of other mobile devices in the same (or adjacent) cell, the wireless network card must be forced to work in promiscuous mode. In this mode, the mobile node consumes more energy.

If the signal strength is sampled at a fade-out, then a handover can be triggered without being necessary. Therefore, the signal is averaged with a time window of a certain size. The antenna system at the receiver as well as at the transmitter impacts the signal strength [8]. To determine L2 trigger, SIR (Signal-to-Interference Ratio), BER (Bit Error Rate) and FER (Frame Error Rate) are mainly used. L2 handover for wireless can be described [9].

(a) MN finds AP1, it will authenticate and associate.
(b) As MN moves, it may pre-authenticate with AP2.
(c) When the association with AP1 is no longer desirable, it may re-associate with AP2.
(d) AP2 notifies AP1 of the new location of the station, terminates the previous association with AP1.
(e) At some point, AP2 may be taken out of service. AP2 would disassociate the associated stations.
(f) MN finds another access point and authenticates and associate.

Now, we present the timeline for L2 handover in Fig. 2. Mishra and Shin [10] found the followings: The probe delay is the dominating component – 90 % of handoff delay. The wireless HW used affects handover latency. There is large variation in handover latency. The different wireless cards follow different sequence. In [11], MN checks whether the corresponding BS (base station) is already in the list of base station within range. If BS is in the list, MN updates the expiration time. Otherwise, MN creates new entry.

Gowasmi [12] proposed that MIPv4 registration message are carried in the Information Elements of 802.11 frame in order to perform fast handoff on the layer 2 for MIPv4.
with 802.11 AP. Yegin [13] proposed that MN sends the reassociation.request to new AP when BER (Bit Error Rate) on the link with the old AP has become too large (contains MAC address of MN). When MN receives the reassociation.reply (contains MAC address of new AP), it trigger IP (mobile-IP) stack. IP stack sends RS to the new AR without waiting RA.

Tan [14] allows AP to advertise the capabilities information of its associated network and deals with IPv6 handover in IEEE802.11. Montavont and Noel [15] found that L2 handover can be very important. When there are several users connected an AP, the L2 handover strongly increase, hence, available throughput is restricted. FMIPv6 offers shorter disruption time than MIPv6. The reason is why MIPv6 is restricted by the time needed to detect the new network prefix. In FMIPv6, tunnel based handover introduces less latency than the anticipated handover because MN does not need to interact with the AR. However, unacceptable delays for real time applications may occur.

In MN movement [16], no all L2 mobility indication from the L2 driver indicates movement of the MN to a new subnet. There are three entities which may change in connection with MN movement, Access Point (Link-layer connection), Access Router and On-link Prefixes (IP Subnet). These changes are indicated to MN with the following:

1. Link-layer triggers
2. A new IP address (in source address field of RA)
3. A new Subnet Prefix (in Prefix Information Option in RA)

To get the above indications, MN can perform ICMPv6 NS (Neighbor Solicitation) /NA (Neighbor Advertisement) exchange, RS/RA exchange or just receive unsolicited RAs.

An MN's movement detection scheme should combine the available information to detect movement correctly. It should not mistake some hint as movement while the MN hasn't moved. That may result in continual handoff, and hence excessive mobility signaling. If the MN moves, it needs to detect movement sufficiently fast so that it can complete handover signaling without significantly degrading application performance. On the other hand, if the MN doesn't move though it receives some hints, it is not imperative to detect its non-movement so fast. It will not degrade performance even if MN can't quickly confirm that it still remains at the same subnet. A movement detection scheme should not result in excessive signaling traffic. It should not flood the network with unnecessary RS/RA or NS/NA messages. The delay time between AP (L2 handover) with the number of users decrease the throughput sharply [16].

3. Layer 3 handover schemes

3.1 Movement Detection
For reachability of old AR and validity of old CoA (Care-of-Address), MIPv4 generally uses LCS (Lazy Cell Switching) based on the <lifetime> field within the ICMP RA, prefix matching (option) based on the prefix-length within the prefix-length extension, and ECS(Eager Cell Switching) based on different network identifier. MIPv6 uses the prefix matching. Perkins [17] insists that for SCTP over ordinary IP v4, we can’t use LCS method since the MN already passes the overlapped region. Especially, advertisement interval time can be set between 3 sec and 30 minutes. This time is so long for fast moving MN. Also, we can’t use the prefix method since the ordinary ICMP router discovery message doesn’t have the prefix or CoA. Thus, we can only consider ECS in ordinary IP v4 environment. In this case, we may use the lifetime or address of router in the RA message.

Trossen [18] assumes that wireless link protocol is capable of delivering a layer 2 identifier for the new AP or the radio interface of new AR to the current AR or to the MN. MN delivers the layer 2 addresses to current AR. Current AR gets the new IP for new AR using layer 2 address.

3.2 Discovery of new AR

Daley [16] uses the timer and NS in order to discover new AR. When MN checks the reachability of current AR, it takes 3 sec. The reason is why MN should send three NS’s during each one second interval. After MN sets timeout, it sends just one NS. If the RA is not arrived, it assumes that it is unreachable.

Choi [19] utilizes fast router discovery with RA caching in AP. AP caches the RA message. AP sends reassociation.reply message and RA to MN simultaneously. MN can configure the IP without sending RS or waiting RA.

Kempf [20] uses IPv6 fast router advertisement. Currently, when the router receives the RS, it should wait random time. Router sends the RA immediately upon receiving RS.

Hong [21] utilizes AR based movement detection and CoA configuration method. AR performs the movement detection. MN sends the L2 trigger to both AP and AR. AR caches MN’s layer 2 addresses and configures the new IP address using the MAC address and its own prefix. When the MN sends the RS, AR sends the RA with the preconfigured CoA.
3.3 New IP address Configuration

Gwon [22] proposed the enhanced forwarding from previous CoA for fast mobile IPv6 handovers. During L2 handoff, MN can find the information of new AR relevant with this ID when it receives the beacon with ID from new AP. Before starting L3 handover, MN can know the information about new AR. In order to do this, MN should maintain the list of candidate ARs using CARID (Candidate Access Router Information Discovery) protocol. When MN makes a bidirectional tunnel to reduce the delay configuring new CoA, it uses the IP address of router as its own temporary IP address. After some time, MN configures its own CoA using the prefix of router.

Moore [23] proposed optimistic duplicate address detection that omits DAD procedure. To reduce the DAD delay, it is proposed to use the tentative IP address. In IPv6, we cannot use the IP address obtained by using stateless-auto-configuration immediately. That is, we can use the valid IP address after performing DAD process. The reason is why there is little probability of conflict. Han [24] proposed the Advance DAT which caches DAD in AR. AR already has the CoA’s pool which has completed the DAD process. When MN sends the modified RS, then AR sends with the modified RA that contains the CoA. MN can use the CoA immediately.

Montavont [15] omitted the DAD or perform it in parallel. Performing DAD generates too much delay in the handover latency. Considering that the probability of address duplication on the same link is extremely low, the MN can choose not to perform DAD or MN should perform DAD in parallel with its communication.

3.4 Binding Updates

MIPv4 first discovers agent. HA and FA inform itself by using AA (Agent Advertisement) message similar to ICMP Router Discovery. MN also can advertise agent using Agent Solicitation message. MN decides if it is in the home network through such an advertisement message.

MN exchanges Registration Request and Registration Reply message with the HA. It registers the CoA in the HA. There are two types of CoA’s: CoA of FA is used for MN (CoA). Temporary IP address by DHCP is used for MN (Co-located CoA). Such a registration message uses UDP port 434 and contains CoA and lifetime of MN.

After successful registration between HA and MN is performed, Datagram sent from CN to the home address of MN is tunneled to the CoA of MN. Datagram from MN to CN has not been tunneled. It is forwarded to the destination using the standard IP routing.

We now investigate Fast Handover (FMIPv6) with MIPv6 which is composed of anticipated Handover and tunnel based handover.

The anticipated handover is performed as followings:

1. MN senses the movement to NAR by using L2 trigger.
2. MN sends PRS (Proxy Router Solicitation) message for NAR to PAR. This message contains link-layer ID for NAR. For example, SSID of NAR in the wireless LAN.
3. PAR configures NCoA using the information of NAR which PAR already has. PAR sends PRA (Proxy Router Advertisement) containing NCoA to MN.
4. MN sends FBU (Fast Binding Update) requesting for binding the old CoA with NAR to PAR (old CoA → NAR).
5. PAR sends HI (Handover Initiate) to setup the bi-directional tunneling with NAR. Also, it requests that NAR will the verification of newly configured CoA.
6. NAR sends HACK (Handover ACK) to PAR. It builds bi-directional tunnel and checks the new CoA.
7. PAR sends acknowledgement for NCoA to NAR through FBACK. It also intercepts data for the previous CoA of MN and forwards to NAR through tunneling.
8. After new link between NAR and MN is established, MN sends RS including FNA (Fast Neighbor Advertisement) which represents MN itself.
9. NAR confirms new CoA to MN through RA with NACCK option.
10. MN sends the BU to both CN and HA.
11. CN/HA reply with ACK.
4. Layer 4 handover schemes

When prefixes of routers are not changed, SCTP (stream control transmission protocol) handover in layer 4 (Fig. 6, [24]) is performed as follows:

1. Physical layer of MN detects the signal power from the new AP when it enters the overlapped region.
2. Physical layer receives beacons/RA from new AP and new AR.
3. MN asks for new IP addresses from DHCP server in new subnet.
4. MN starts handover with IP addresses.
5. MN sends ASCONF:add_ip (new) chuck to CN.
7. MN sends ASCONF: set_primary_IP(new) to CN.
8. CN replies ASCONF-ACK that primary has been switched.
9. MN sends ASCONF:delete_IP(old) to CN.
10. CN replies ASCONF-ACK:delete_IP(old) to MN.

In step (3), it is reasonable that first, to obtain the IPv6 address first by using stateless-auto-configuration, if we fail, then use the state full method such as DHCP. The reason to do this is that if the DHCP server is far away from the location of MN, the time to send the DHCP request message and receive the DHCP reply can be longer than the time to obtain the prefix from Access Router attached to MN. That is, as soon as we receive the RA from the attached router prior to receiving the RA, we...
can send the RS. Hence, we obtain the IP-address by using the stateless-auto-configuration of IPv6 (MIPv6).

![Fig. 6 Timeline for L4 (SCTP) handover](image)

If this procedure fails, we can use the state full method such as DHCP based on MAC address. Assume that we use the FMIPv6 (Fast Handover for Mobile IPv6) when we detect the signal from the physical layer. Then, we can get the CoA of new router (subnet B) prior to the handover. Therefore, we can find new IP address of host by using the stateless-auto-configuration.

5. Conclusions

In this paper, we address the layer based handover schemes. Layer 2 handover mainly depends on how to detect the signal power and is performed using hardware. The performance of Layer 3 handover scheme is affected by how to exchange old IP address with new IP address rapidly. Additionally, how to obtain new IP address fast is the important factor for handover latency. Generally, MIPv4 and MIPv6 including FMIPv6 are used for supporting mobility. L4 handover scheme using SCTP is very simple and known to have better performance than TCP based handover schemes. However, it has cross-layer problem and requires the dual interfaces. Efficient handover mechanism is one of the important issues in mobile network in order to obtain more fast communication. This paper describes the pros and cons of several handover schemes, which help us to develop new handover mechanism.

References


Mobility analysis and Framework proposal on Clustering and Validation Mechanisms for Ad Hoc Network

S. Thirumurugan¹, Dr. E. George Dharma Prakash Raj²

¹ Research Scholar, Computer Science Department, Bharathidasan University
Trichy, Tamilnadu, India
Email s.thiru.gan@gmail.com

² Computer Science Department, Bharathidasan University
Trichy, Tamilnadu, India
Email georgeprakashraj@yahoo.com

Abstract
The mobility of the nodes affects the compactness of the clusters. This ends up with the decline in the performance of the clustered ad hoc networks. Thus, while the clusters move to a worst state in their performance, re-clustering should be suggested. But this process will be considered as overhead of the network if the re-clustering time hasn’t been decided properly. The validation mechanisms should be applied to decide the re-clustering time of the ad hoc network. This research work shows the analysis of the clusters periodically where the nodes are dynamic in nature. This analysis has been done for 10 nodes as a sample set. The measures are done for time periods T₁, T₂, T₃ and T₄. The results are tabulated to make a decision on the re-clustering. This work also incorporates a framework for the clustering and validation mechanism.

Keywords: DB Index, Dunn’s Index, Silhouette Index, W-PAC.

1. Introduction

Over the decades, the communication system has grown tremendously in an unprecedented manner. The wired mode can no longer dominate the communication system. The wireless mode comes here not merely as a supportive instead this has begun its revolution to replace the wired communication system. This trend stands as a fair and cost-effective approach. The wireless network refers to any type of computer network that uses wireless (usually, but not always radio waves) for network connections. It is a method by which homes, telecommunication networks and enterprise (business) installations avoid the costly process of introducing cables into a building, or as a connection between various equipment locations. Wireless telecommunication networks are generally implemented and administered using radio communication. This implementation takes place at the physical level of the OSI model [1] network structure. This wireless networks may or may not be based on the infrastructure. Wi-Fi with access points and mobile phone communication with base stations are some of the examples for infrastructure based networks.

The field of wireless and mobile communications has experienced an extraordinary growth during the past decade. Current Second-Generation(2G) cellular systems have reached a high penetration rate enabling worldwide mobile connectivity. Mobile users can use their cellular phone to check their email and browse the internet. Recently, an increasing number of wireless local area network hot spots are emerging, allowing travelers with portable computers to surf the internet from airports, railways, hotels and other public locations. However, all these networks are conventional wireless networks, conventional in the sense that as prerequisites, a fixed network infrastructure with centralized administration is required for their operation, potentially consuming lot of time and money for set-up and maintenance. These limitations made us to understand the need of the ad hoc networks [2].

The performance of ad hoc networks could be maintained to some extent while the nodes are on move. There should be some mechanism to improve this situation. The clustering is one such mechanism to ensure frequency reusability and handling modular growth. There are various clustering algorithms [3][4] proposed to handle this situation. Apart from clustering, validation has been suggested to decide the re-clustering time.

2. Literature Review

In the literature, many authors have proposed single and multi-parametric cluster creation algorithms. Lowest ID [5], Highest Degree [6] [7] and k-means [8] are proposed under single parameter category. K-means has been accepted as a well-known single parameter based algorithm. In the case of multi-parametric category several
algorithms are proposed. They are WCA [9], EWCA [10], EBC [11], WBCA [12] and FWCA [13]. WCA has been considered as well accepted multi-parametric algorithm.

The validation of clusters has been done by validation indices namely silhouette index [14], Dunn’s index and Davies-Bouldin index [15]. Based on the existing research works, the following drawbacks are observed. They are: (i) Single and Multi-parametric algorithms are taking more time to form the clusters. (ii) There is no procedure to test the strength of the clusters formed. (iii) These algorithms lack in specifying the sustainability of the clusters which obviously determines the re-clustering time of ad hoc network. To overcome those aforementioned drawbacks, the cluster formation methods and validation techniques have been proposed. The mobility analysis which is making use of clustering and validation mechanism helps us to decide the re-clustering time of ad hoc networks.

3. Clustering and Validation Indices

In this study the clusters have been created using multi-parametric algorithm (W-PAC) [16]. The validation on the clusters has been done using the following indices.

Davies –Bouldin Index

\[ \frac{1}{n} \sum_{i=1}^{n} \max_{l \neq j} \left( \frac{C_i(N_l) + C_j(N_l)}{C(N_l,N_j)} \right) \]

Dunn’s Index

\[ \min_{p=1...n} \min_{q=p+1...n} \frac{\text{dis}(C_p,C_q)}{\max_{r=1...n} \text{dia}(C_r)} \]

Silhouette Index

\[ S(i) = \frac{\text{b}(i) - \text{a}(i)}{\max[\text{a}(i), \text{b}(i)]} \]

4. Mobility analysis

This analysis has been divided into several stages with respect to time factor. The results obtained based on this analysis decides the re-clustering time of the ad hoc networks to sustain the performance of the network.

**Time T1:** Time T1 represents the state immediately after the clusters have been created using multi-parametric (W-PAC) algorithm.

**Time T2:** Time T2 represents the structure of the ad hoc network after the node N5 makes a move from Cluster C1 to Cluster C2.
Figure 3 shows the clusters and member nodes at time T2. These results are obtained after the mobility of node N5 from Cluster C1 to Cluster C2.

Figure 4 shows the graphical results of validation at time T2. This obviously tells that the value of DB index and Dunn’s index remain same as T1. Silhouette index value has down by 2%. This reduction indicates the decline in compactness of the clusters which means the member nodes are moving away from clusterheads. This reduces strength of communication signal between members and their respective clusterhead.

**Time T3:** At this time node N1 has moved from cluster C1 to cluster C2. This mobility further reduced the size of the cluster C1 and increased the size of cluster C2. The clusterhead of cluster C1 handles less number of nodes. As a result of this, the burden of clusterhead of C1 gets down. This is because of very few number of nodes initiates the data transfer operation with the clusterhead of cluster C1. In the case of Cluster C2, the number of data transfers increases proportionately as the number of affiliation of nodes increases. The threshold may limit the number of nodes of clusterhead. This situation makes the clusters to reach imbalance state.

Figure 5 shows the mobility of the node N1 from cluster C1 to C2 at time T3. This mobile node N1 gets affiliated under the new clusterhead by sending the ‘hello’ message to new clusterhead node.

Figure 6 shows the graphical representation of the measured validation indices at time T3. These values show a decline in silhouette index value while this is compared with time period T2. This silhouette index value when it goes down below the threshold the sustainability of the cluster reduces.

**Time T4:** At this time period T4, the node N2 makes a move away from clusterhead of the cluster C1. As a result of this, the distance between node N2 and clusterhead of the cluster C1 has increased. The communication between the clusterhead and member node N2 breaks since N2 has crossed the boundary of cluster C1.
Figure 7 Results of Clusters at time T4

Figure 7 shows the result of clusters at time T4 after the mobility of node N2. The cluster C1 has less number of nodes whereas C2 has more number of nodes.

Table 1 Consolidated results of cluster validation at time T4

<table>
<thead>
<tr>
<th>Validation Index</th>
<th>Nodes</th>
<th>Index Value</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB Index</td>
<td>10</td>
<td>0.944</td>
<td>94.4%</td>
</tr>
<tr>
<td>Dunn's Index</td>
<td>10</td>
<td>0.57</td>
<td>57%</td>
</tr>
<tr>
<td>Avg. Silhouette Index</td>
<td>10</td>
<td>0.47</td>
<td>47%</td>
</tr>
</tbody>
</table>

Table 2 Cluster validation results of 10 Nodes

<table>
<thead>
<tr>
<th>Time</th>
<th>DB Index</th>
<th>Dunn's Index</th>
<th>Avg. Silhouette Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>T₁</td>
<td>0.67</td>
<td>1.0</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>67%</td>
<td>100%</td>
<td>63%</td>
</tr>
<tr>
<td>T₂</td>
<td>0.67</td>
<td>1.0</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>67%</td>
<td>100%</td>
<td>63%</td>
</tr>
<tr>
<td>T₃</td>
<td>0.694</td>
<td>0.8</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>69.4%</td>
<td>80%</td>
<td>60%</td>
</tr>
<tr>
<td>T₄</td>
<td>0.944</td>
<td>0.57</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>94.4%</td>
<td>57%</td>
<td>47%</td>
</tr>
</tbody>
</table>

Table.1 shows the consolidated results of cluster validation at time T₄. It shows an increase in DB index value, decrease in Dunn’s index value and decrease in Silhouette index value while this is related with the values measured at time T₃.

Table.2 shows the cluster validation results of 10 nodes as sample size. It shows the consolidated results of the measure of hard validation indices at various time periods T₁, T₂, T₃ and T₄. These results need to be observed in order to conclude the sustainability of the cluster based on threshold fixed by the application for which the clustered ad hoc network has to be formed.

Figure 8 Results of cluster validation from T₁ to T₄.

Figure 8 shows the results by graphical representation. The following results are observed at various time periods T₁, T₂, T₃ and T₄.

At time T₁, the results show that the clusters are in good condition. They are well compact in nature.

At time T₂, the results show that the compactness sustains without noticeable change due to the movement of the node N₅. Silhouette index has got reduced slightly to indicate a negligible amount of reduction in clusters compactness.

At time T₃, the DB index value has increased to indicate the fall in the strength of intra-cluster relationship. This rise in DB index proportionately increases the intra-cluster distances. The Dunn’s index value has reduced by 20% while it is compared with the previous value. This obviously shows the increase in overlap level of the clusters. Thus, compactness of clusters which is shown by silhouette index has reduced.

At time T₄, the rise in DB index and fall in Dunn’s index value further shows the weak state of the clusters formed.
using the cluster formation algorithms. This would be considered as the re-clustering time since the values are going down. This may also drive the re-election of clusterhead in the Cluster $C_2$ since the number of nodes has increased, which will exhaust the energy level of clusterhead.

This research work has emphasized on reducing the overhead of the network. The re-clustering has been procrastinated till $T_4$ since the re-clustering should not be done for a slight change in the structure. As the re-clustering process has been considered as overhead of the networks, this should be postponed to the maximum extent possible.

5. GDPRS-C&V Framework

The following scenarios are considered to implement this cluster formation and validation.

- To set up conference network within the Hall the inter cluster distance may likely to be maximum. Since the network will be constant for a while needs smooth communication across the clusters of the network.
- To set up field study(soil humidity) network the inter cluster distance should be as maximum as possible. Since this network would be fixed for the specific period.
- To set up network within the campus where the nodes are highly mobile deserves less inter cluster distances. This is to keep the re-clustering process to happen while it is desperately needed.

![Diagram](Figure 9 GDPRS – C&V : Framework for Ad Hoc Networks)
The framework for ad hoc networks has been devised by Dr. George Dharma Prakash Raj and Mr. Shomasundaram Thirumurugan (GDPRS); C & V stands for Clustering and Validation.

Figure 9 shows the Clustering and Validation Framework for Ad-Hoc networks.

The cluster creation is the first phase of this framework. It tells about the options of cluster creating algorithms. These algorithms are divided as single-parametric and multi-parametric algorithm. The selection of algorithm depends on the real world scenario for which ad hoc network has to be set up.

The second phase of framework is covered by cluster validation process. The clusters which are created using cluster formation algorithms will reach this second phase as input to be processed further. This second phase just checks the cluster status to make some decision on those measurements. This second phase is subdivided into parts such as hard indices based validation and soft fuzzy logic based validation.

The hard validation indices periodically measure the cluster perfection or compactness. It recommends for re-clustering while the validation measures R (Dunn’s, Davies-Bouldin and Silhouette index) goes down below the TH (threshold) which depends on the specific application.

The soft fuzzy logic based validation procedure helps to confirm the elected clusterhead of the clusters of ad hoc network. In case of the clusterhead mobility this approach re-elects the clusterhead. This also procrastinates the re-clustering process by re-election of clusterhead locally. This also helps to elect the gateway to ensure inter-cluster communication among the clusters of the ad hoc network.

6. Conclusion

This study helps to decide the re-clustering time of the ad hoc networks. This has been decided with the help of clustering and validation on the mobility scenario of the network. The findings will ensure the sustainability of the network to strengthen the application for which the network has been utilized. A framework proposal aids on understanding the complete process of clustering and validation.

References


Mr. S. Thirumurugan completed his Masters Degree in Computer Applications and Master of Philosophy in Computer Science. He has around 10 yrs of experience in teaching field which includes his association with the research work. He has published his work in six international journals, presented four papers at the international level and also two papers at the national level. His area of research work falls on Ad hoc networks and their applications on real world scenario.
Dr. E. George Dharma Prakash Raj completed his Masters Degree in Computer Science and Master of Philosophy in Computer Science in the years 1990 and 1998. He has also completed his Doctorate in Computer Science in the year 2008. He has more than twenty years of Academic experience and thirteen years of Research experience in the field of Computer Science. Currently he is working as an Assistant Professor in the Department of Computer Science and Engineering at Bharathidasan University, Trichy, India. He has published several papers in International Journals and Conferences related to Computer Science and has been an Editorial Board Member, Reviewer and International Programme Committee Member in many International Journals and Conferences.
A survey on web penetration test

Mahin Mirjalili¹, Alireza Nowroozi², Mitra Alidoosti³

¹Department of information security, Malek-Ashtar university of technology, Tehran, Iran
Mahinmirjalili@mut.ac.ir

²Department of information security, Malek-Ashtar university of technology, Tehran, Iran
Nowroozi@mut.ac.ir

³Department of information security, Malek-Ashtar university of technology, Tehran, Iran
alidoosti@mut.ac.ir

Abstract

This paper reviews the penetration test specifically in the field of web. For this purpose, it first reviews articles generally on penetration test and its associated methods. Then articles in the field of web penetration test are examined in three aspects: comparing automatic penetration test tools, introduction of new methods or tools for manual penetration test, and articles that presented a test environment for training or checking various instruments and methods. This article studied 4 different methodologies for web penetration test, 13 articles for comparing web vulnerability scanners, 10 articles that proposed a new method or tool for penetration test and 4 test environments.

Keywords: Penetration test, web scanner, web application, web vulnerabilities.

1. Introduction

Penetration test is a security evaluation process for network or computer systems that simulates an attack by an ethical hacker. The most important distinction between a hacker and a penetration tester is that penetration test is done with a license and a signed contract with an organization or company, and the output is provided as a report. The goal of penetration test is to increase data security. Security information and weaknesses that are specified in penetration test are considered confidential and shall not be disclosed until complete resolution of defects.

Given the importance of web application security, this article reviews studies in the field of penetration test, particularly web penetration test.

Questions that engaged the mind of researchers in the field of penetration test can be expressed as follows:

- How the penetration test is performed?
- What are types of penetration test?
- How the penetration test is done automatically?
- What tools can we use to perform an automatic penetration test?
- Comparison of tools and their effectiveness
- What are the new tools and methods and what are their features?
- How can we examine various tools and techniques?

This paper attempts to answer these questions by examining 4 different methodologies for penetration test, 13 articles for comparing web vulnerability scanners, 10 articles that proposed a new method or tool for penetration test and 4 test environments.

Due to the large volume of papers in the studied area, some criteria were used for paper selection such as covering a wide time period from 2006 to 2014 and the number of citations per paper. Most selected articles were considered by many authors in previous years.

We will initially review articles that provided a method for the penetration test, then papers in the field of web penetration test in three views:

- Articles on the comparison of existing methods and tools for web penetration test.
- Articles that proposed a new tool or method for web penetration test.
• Studies that proposed a test environment for testing tools and ideas.

2. Penetration testing

2.1. History and Importance

Penetration testing is one of the oldest methods for assessing the security of a computer system. In the early 1970’s, the Department of Defense used this method to demonstrate the security weaknesses in computer systems and to initiate the development of programs to create more secure systems. Penetration testing is increasingly used by organizations to assure the security of Information systems and services, so that security weaknesses can be fixed before they get exposed [1].

Large companies have important data and one of their concerns is to protect the data. The penetration test tests security mechanisms of companies by simulating multiple attacks. In the situations like new infrastructure is added, Software is installed, System updates are applied, Security patches are applied and User policies are modified it is necessary to do penetration testing. Some of the principal reasons for adopting penetration testing are Security Issues, Protect Information, Prioritize security risks and Financial Loss [2].

The penetration test can be done either manually or automatically. Manual penetration test requires a skilled and experienced tester team to control all things. They must be physically present at the test duration. Thus this is not an affordable option. Automatic penetration test is a simple and safe way to perform all tasks related to the penetration test. Moreover, since most work is done automatically, it is more economical in terms of time. Another advantage of this test is the ability to reuse the set parameters for the test. In [3], a comparison is made between the two tests, as follows:

<table>
<thead>
<tr>
<th>Manual</th>
<th>Automated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testing Process</td>
<td>Labor-intensive, inconsistent and error-prone, with no specific quality standards. Requires many disparate tools. Results can vary significantly from test to test. Generally requires highly-paid, experienced security personnel to run and interpret tests.</td>
</tr>
<tr>
<td>Network Modification</td>
<td>Often results in numerous systems modifications.</td>
</tr>
<tr>
<td>Exploit Development and Management</td>
<td>Developing and maintaining an exploit database is time-consuming and requires significant expertise. Public exploits are suspect and can be unsafe to run. Re-writing and porting code is necessary for cross-platform functionality.</td>
</tr>
<tr>
<td>Cleanup</td>
<td>Tester must remember and undo all changes. Backdoors can be left behind</td>
</tr>
<tr>
<td>Reporting</td>
<td>Requires significant effort, recording and collating of all results manually. All reports must be generated by hand.</td>
</tr>
<tr>
<td>Logging/ Auditing</td>
<td>Slow, cumbersome, often inaccurate process.</td>
</tr>
<tr>
<td>Training</td>
<td>Testers need to learn non-standardized, ad-hoc testing methods.</td>
</tr>
</tbody>
</table>
Penetration testing can be segregated into the following classes [4]: Attack visibility: Blue-teaming or Red-teaming, and system access: Internal testing or External testing.

Blue-teaming is done with the consent of an entire organization. The information security team is fully aware of the testing requirements as well as resources needed. Blue-teaming is a more efficient way to perform testing as the system availability is not an issue and hence there is a considerable reduction in the overall time for testing. The shorter test times mean lesser system idle time and reduced testing costs. Red-teaming refers to testing that is performed in a stealth manner without the knowledge of IT staff [4]. Upper-level management authorizes such an exercise. The objectives of the test are to judge the strength of the network security, the awareness of IT organization, and its ability to follow the standard protocols. The entire test is done without the support of the organization’s resources.

2.2. Available methodologies and techniques

A methodology is a scheme that is used to reach the destination. Lack of use of a methodology for penetration test may lead to an incomplete test, high time-consumption, failure and test ineffectiveness. Despite the large number of methodologies, there is nothing called "true methodology" and each penetration test can have a different methodology but the use of a methodology leads to a professional and efficient penetration test at lower cost.

The methodology considered for the penetration test usually has 4 or 7 phases. Although the name or number of phases is different in different methodologies, they all show an overview of penetration test. For example, some methodologies use the term "information gathering" while some call this process "reconnaissance."

The methodology proposed in [5] consists of four phases: reconnaissance, scanning (port scanning, vulnerability scanning), exploitation and maintaining access. The first phase of the penetration test is reconnaissance which focuses on information gathering. The more information gathered at this stage, the more successful the next stages will be. The second phase of this methodology is divided into two categories: port scanning which obtains a list of open ports and services running on each of them, and vulnerability scanning which is a process to recognize weaknesses of desired services and applications. According to the results obtained in step 2 and knowing what ports are open, what services are running on this port and what vulnerabilities they have, one can attack the target. Maintaining access is the last phase. Most accesses obtained in the attack phase are temporary, and are removed after disconnection. In this phase, it is tried to maintain access. Although this reference ignored "reporting" as a step in the penetration test, it stated the last activity of a penetration test as reporting. According to [5], reporting should include details on how to perform the test, a summary of the found security threats, cases the test does not cover, etc.

![Figure 1 - proposed methodology in [5]](image)

In the NIST penetration test methodology, the penetration test consists of four phases: planning, discovery, attack, reporting. In the planning phase, rules are defined and objectives of the test are set. The discovery phase is performed in two stages. The first includes test initiation and information collection and the second stage, which takes place after the attack phase, includes vulnerability analysis. In the attack phase, which is known as the heart of penetration test, various vulnerabilities in the target are examined. The report is prepared in conjunction with the other phases. In the planning phase, the evaluation plan is developed. In the discovery and attack phases, events are usually recorded and periodically reported to the director. At the end of the test, a report is provided to describe recognized vulnerabilities, ranking of risks, and tips for how to improve the known weaknesses [6].
The methodology presented in [7] consists of three main parts: information, team, and tools. In the information part, information is gathered on the target using different methods. In this paper, the information phase is defined in four steps: studying the network, identifying the OS, scanning ports and identifying services. The second part of the methodology is team formation. If teams are formed with different roles and responsibilities, the penetration test will be carried out more effectively. Another important parameter in the penetration test is to use tools. To do an effective penetration test, it is better to dominate a smaller number of tools, instead of many tools. Another point discussed in this article is to set policies that must be followed by the tester and the client. In the proposed policy, there are issues such as preservation of information obtained by the tester and reporting them completely at the end of the penetration test, scheduling agreement, confidentiality of all information such as contract, use of information obtained just for the test, lack of responsibility of the penetration tester in the event of a real attack, and so on.

In the induction phase, the time period and type of the test should be specified. The interaction phase indicates the objectives of the penetration test. In the inquest phase, the maximum possible data is achieved on the target system. In the final phase, the security performance is measured. After completion of the penetration test, results are processed and a report is prepared. The OSSTMM uses a set of tools called Security Test Audit and Reporting for processing the results [8].

In [41], a penetration test scheme is presented for web application based on the RUP test scheme. Each of the described methodologies may be appropriate for different purposes and penetration tests. As mentioned earlier, it cannot be said that a methodology is better than another. This scheme provides a systematic, consistent and affordable method fully integrated with the security-based software development lifecycle for the penetration test and improves the accuracy, quality and performance of such tests. This study also presents a database of techniques and tools required for the penetration test of web applications which was compiled using various sources including valid guidelines and standards and test techniques on the Internet.

In [42], an penetration test methodology is presented based on the agile method which uses the benefits of the agile method in the process of penetration test, and a model is design based on the Scrum and XP methodologies which show information flows between activities. Thus, this method improves the penetration test cycle and can be a framework for improving the accuracy, efficiency, job satisfaction and quality of testers. In this process, change management can be easily done and information technology goals are aligned with business goals, interaction with customer increases, so prioritizing the depth and range of penetration test is easier.

3. Web penetration test

Today, with the Internet expansion and use of web applications in various fields such as military, medical, finance, etc. web security is an important concern, and the penetration test is used to ensure it. The penetration test can be performed manually or
automatically. The two options are compared in Table 1.

Tools used to recognize vulnerabilities can be divided into three categories based on information of the target application that they use: white-box, black-box and grey-box.

A white-box tool uses the target application code to assess vulnerabilities. By analyzing the code application, a white-box tool can find all the hidden application paths which lead to finding vulnerabilities in the application path. In this set of tools, due to access to the application code, vulnerabilities may be reported that are not available. In other words, it is likely that there is no possibility to use a known vulnerability. The disadvantage of white-box test is its reliance on a specific language and framework.

Unlike white-box tools, black-box tools assume that there is no knowledge of the application code. Instead of using the application code, the tester, like a regular user, uses the application with a browser. A black-box tool first examines different parts of the application to find all possible injection vectors. Whatever way through which the attacker can enter the application is called attack vector like URL parameters, HTML form parameters, cookies, HTTP headers etc. After identifying injection vectors to the application, inputs are given to recognize vulnerabilities. This process is called fuzzing. In fuzzing, the type of injection vector and its use duration can be different in black-box tools. Finally, these tools examine http and html responses related to fuzzing, and if successful, report it as vulnerability.

Among advantages of white-box over black-box tools, one can point to independence of the application code and less false positive. Given that a black-box tool can only identify vulnerabilities that run their related attack, one of its disadvantages is that it cannot guarantee to identify all the vulnerabilities of the application.

Grey-box tools, as their name implies, are a combination of black-box and white-box. These tools use static white-box analysis techniques to identify vulnerabilities. Then they try to really attack identified vulnerabilities to confirm them. If this step is successful, the vulnerability is reported. Grey-box tools can find vulnerabilities in all the application paths with low false positive but, like white-box tools, they depend on a specific language or framework [9].

3.1. Web vulnerabilities

Given that applications are highly vulnerable, invaders use different methods and paths to damage various organizations. These vulnerabilities can be very simple or complex. In complex ones, the discovery and exploitation become very difficult for invaders. According to the characteristics of its business environment and associated risks, especially threatening factors, each organization identifies implemented security controls and their impact on business and financial matters of the organization. Owasp annually publishes a list of ten common vulnerabilities. The last list published in 2013 includes the following vulnerabilities [36]:

1- Injection: Injection flaws, such as SQL, OS, and LDAP injection occur when untrusted data is sent to an interpreter as part of a command or query. The attacker’s hostile data can trick the interpreter into executing unintended commands or accessing data without proper authorization.

2- Broken Authentication and Session Management: Application functions related to authentication and session management are often not implemented correctly, allowing attackers to compromise passwords, keys, or session tokens, or to exploit other implementation flaws to assume other users’ identities.

3- Cross-Site Scripting (XSS): XSS flaws occur whenever an application takes untrusted data and sends it to a web browser without proper validation or escaping. XSS allows attackers to execute scripts in the victim’s browser which can hijack user sessions, deface web sites, or redirect the user to malicious sites.

4- Insecure Direct Object References: A direct object reference occurs when a developer exposes a reference to an internal implementation object, such as a file, directory, or database key. Without an access control check or other protection, attackers can manipulate these references to access unauthorized data.

5- Security Misconfiguration: Good security requires having a secure configuration defined and deployed for the application, frameworks, application server, web server, database server,
and platform. Secure settings should be defined, implemented, and maintained, as defaults are often insecure. Additionally, software should be kept up to date.

6- Sensitive Data Exposure: Many web applications do not properly protect sensitive data, such as credit cards, tax IDs, and authentication credentials. Attackers may steal or modify such weakly protected data to conduct credit card fraud, identity theft, or other crimes. Sensitive data deserves extra protection such as encryption at rest or in transit, as well as special precautions when exchanged with the browser.

7- Missing Function Level Access Control: Most web applications verify function level access rights before making that functionality visible in the UI. However, applications need to perform the same access control checks on the server when each function is accessed. If requests are not verified, attackers will be able to forge requests in order to access functionality without proper authorization.

8- Cross-Site Request Forgery (CSRF): A CSRF attack forces a logged-on victim’s browser to send a forged HTTP request, including the victim’s session cookie and any other automatically included authentication information, to a vulnerable web application. This allows the attacker to force the victim’s browser to generate requests the vulnerable application thinks are legitimate requests from the victim.

9- Using Components with Known Vulnerabilities: Components, such as libraries, frameworks, and other software modules, almost always run with full privileges. If a vulnerable component is exploited, such an attack can facilitate serious data loss or server takeover. Applications using components with known vulnerabilities may undermine application defenses and enable a range of possible attacks and impacts.

10- Invalidated Redirects and Forwards: Web applications frequently redirect and forward users to other pages and websites, and use untrusted data to determine the destination pages. Without proper validation, attackers can redirect victims to phishing or malware sites, or use forwards to access unauthorized pages.

3.2. Black-box web vulnerability scanners

In recent years, automated or semi-automated application scanning has been addressed to find vulnerabilities. To perform an automatic penetration test, there are many commercial and open source scanners. Next, we introduce some commercial and open source scanners and then review research on the web penetration test. Articles in this category have considered three issues: comparing scanner available tools, developing a new scanner and designing a vulnerability application.

3.2.1. Open-source vulnerability scanners

Table 2 presents some open source scanners and their developer companies, used technology and platform. Then Table 3 compares them in terms of use scale, initial scan method and outputs of each one.

<table>
<thead>
<tr>
<th>Table 2-Open-source scanners overview [39]</th>
</tr>
</thead>
<tbody>
<tr>
<td>scanner</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
</tbody>
</table>

Copyright (c) 2014 Advances in Computer Science: an International Journal. All Rights Reserved.
3.2.2. Commercial vulnerability scanners

Commercial vulnerability scanners are developed by various organizations. Although they cost too much, they have fewer security bugs compared to open source scanners. Because of competition between different organizations for developing scanners, their limitations are not disclosed. Table 4 introduces some popular commercial scanners.

3.3. Academic research about web scanners

Articles in the field of penetration testing are analyzed in three groups. A group of articles compared commercial and open source scanners and tested them against some vulnerability. A second group is articles that provide a new method or tool for automatic penetration testing. Moreover, for assessment of security methods and tools, we require vulnerable applications whose vulnerabilities are clear. The third group of studies is devoted to these applications.

3.3.1. Comparison of existing tools and methods

Numerous articles reviewed and compared web scanners. In some of these papers, the comparison of tools is the main goal, and some others compared tools with the goal of providing a new tool or technique. Most vulnerabilities considered in these comparisons are SQL and XSS injection. Many of these articles suggest that available tools cannot identify all vulnerabilities and some others concluded that false positive is high.
Table 2 and Table 3 present the scanners tested in each article, vulnerabilities and the test environment.

In [10], the name of tools has not been mentioned for commercial reasons and the neutrality of the article.

In [11], results indicate that different tools produce different results, cannot identify many vulnerabilities and have about 20% to 70% false positive.

The comparison by McAllister et al. (2008) shows that tools have a low ability to detect vulnerability, however, using the proposed technique, more vulnerabilities were found [12].

In [13], the authors, due to the detection rate of vulnerabilities, claimed that their test suit is effective for different tools. They also stated that none of the tools are capable of identifying level 2 or higher vulnerabilities.

Shelly (2010) concluded that the use of a test environment with secure and insecure versions is a good way to find the reasons of producing false positive and false negative by tools. The conclusion declared on the quality of tools indicates that tools can detect simple XSS and SQL injection. But to identify non-simple XSS and SQL injection, session management flows, running malicious files and buffer overflow, more work is needed to improve techniques and tools [14].

In [15], it is noted that tools require a greater understanding of active contents and scripting languages like SilverLight, Flash, Java Applet and JavaScript.

In [16], a new tool called CIVS-WS9 is developed with a new method to identify SQL/XPath injection. It came to the conclusion that the implemented tool has the 100% coverage power and 0% false positive.

The paper [17] is based on the results obtained in [10]. The author proposed a method to identify SQL injection and developed a tool called VS.WS. To test this method, test [10] was repeated. All tools were executed against 262 public web services and Java implementation from 4 web services specified By TPC-APP benchmark. It concluded that the implemented tool performance in terms of coverage and false positive is better than commercial tools.

In [18], it is concluded that crawling in an advanced web application is a serious challenge for penetration test tools and they require supporting technologies such as Flash and Java, more advanced algorithms for performing deep crawling and tracking the state of the application under test and more studies on automating the identification of vulnerabilities in the application logic.

In [19], the authors concluded that even when scanners are taught to exploit vulnerabilities, they cannot detect stored SQL injection. They also state that improving some of the functionality of black-box scanners like state full scanning, input selection by field name and tag, the novelty of attack vector, server response analysis and post scanning can improve the discovery rate.

Reference [20], which is a generalization of [15, 18], studied three scanners and concluded that scanners cannot detect vulnerabilities because of weaknesses in the third phase.

In [21], it is concluded that Iron WASP, NetSparker community edition, OWASP ZAP, Vega, N-Stalker and W3AF identified highest to lowest number of vulnerabilities and had the lowest to highest false negative, respectively.

Based on the results of [22], W3AF, Archani and Skipfish were chosen as the best. It was also shown that the most difference in scanners is in identifying injection vulnerabilities, cross-site scripting, session management and broken authentication.

<table>
<thead>
<tr>
<th>Table 2-Scanners evaluation summery</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scanners</strong></td>
</tr>
<tr>
<td>[10] HP WebInspect, IBM Rational AppScan, Acunetix Web Vulnerability Scanner</td>
</tr>
<tr>
<td>[12]</td>
</tr>
</tbody>
</table>

9 Command Injection Vulnerability Scanner for Web Services
In this section, we reviewed several papers that compared different scanners. As seen in Table 7, vulnerabilities addressed in Owasp Top 10 were considered in these papers. To enhance the value of assessments, most articles used popular scanners Like Acunetix Web Vulnerability Scanner, IBM Rational AppScan, and Burp scanner. Table 8 summarizes the number of using each scanner in articles.

Table 4 - Frequency of used scanners in papers

<table>
<thead>
<tr>
<th>Scanners</th>
<th>Used in papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Acunetix Web Vulnerability Scanner</td>
<td>8</td>
</tr>
<tr>
<td>(2) IBM Rational AppScan</td>
<td>6</td>
</tr>
<tr>
<td>(3) w3af, (4) N-stalker</td>
<td>5</td>
</tr>
<tr>
<td>HP WebInspect, Brup Spider, Grendel-Scan,</td>
<td>3</td>
</tr>
<tr>
<td>Hailstorm, Wapiti, OWASP ZAP</td>
<td></td>
</tr>
<tr>
<td>Netsparker, VS.BB, Paros, Iron WASP, Vega</td>
<td>2</td>
</tr>
<tr>
<td>other</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5 lists attacks to different vulnerabilities and Table 8 lists scanners that were investigated in the compared articles more than others and shows that they cover what percentage of cases in Table 6(those not listed in the table are covered by all scanners) [37].

Table 5 - Attacks

<table>
<thead>
<tr>
<th>#</th>
<th>Attack</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Error based SQL injection</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>Blind SQL injection</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>Server Side Java Script injection</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>Reflected Cross Site Scripting</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>Persistent Cross Site Scripting</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>DOM Cross Site Scripting</td>
<td>23</td>
</tr>
</tbody>
</table>

10 Command Injection Vulnerability Scanner for Web Services
In Simple PHP Blog, Serendipity and 36 known vulnerabilities were recreated with 27 false positives. In Simple PHP Blog, Serendipity and 36 known vulnerabilities were recreated with 27 false positives. A restriction of Pixy is that it does not support 'OO' (object orientation).

Saner is a tool developed by combining static and dynamic techniques to identify errors in applications in PHP language [24]. It uses static techniques to identify inputs in each path with the help of string modeling methods. The second technique is dynamic analysis which operates bottom-up. It first recreates codes to identify inputs and then uses a large collection of malicious inputs to identify exploitable flows. The dynamic phase aims to test all the application paths that were identified suspicious in the static phase. Since the static analysis phase is prudent, it is possible to produce false positive which must be manually evaluated which is a boring process. The purpose of automatic dynamic analysis is to automate this process or at least automate detection of what vulnerability input should be used. For assessment, this tool was evaluated on five applications of Jetbox, MyEasyMarket, PBLGuestbook, PHP-Fusion and Sendcard which are developed in PHP. Although this tool has not a very good execution performance, the time required to analyze most applications is good.

In [25], the most common input validation vulnerability model called Tainted Mode Model is used to identify internal vulnerabilities. This paper improves the classical tainted mode model to investigate internal data flows. It also introduces a new method using information obtained from dynamic analysis for doing the automated penetration test. Given a larger view it gives from the application, its accuracy will be more, and the accuracy of input validation procedures can be tested. Using improved tainted mode model, applications were modeled as follows:

\[ W: \text{Scheme, Req} \times \text{State} \rightarrow \text{DDG} \times \text{Resp} \times \{\text{query}\} \times \text{State} \]

Where scheme is a set of relational data schemes which shows the application database. Req is the http request sent to the application. State, on the left, refers to the application state which includes contents of the application environment (such as a database, system files and LDAP). DDG = (V, E) is a data

Since XSS and SQL injection vulnerabilities are among common vulnerabilities, they were reviewed in the majority of articles.

### 3.3.2. New methods and tools

This section reviews a number of papers that designed a new scanner.

In [23], a tool was design with static analysis capabilities called Pixy to identify vulnerabilities in web applications. Pixy is an open source tool and its main purpose is to identify cross-site scripting vulnerability in PHP scripts. PHP was selected because it is widely used in designing web applications and by a large number of security consultants in PHP applications. Vulnerability detection was based on data flow analysis. To evaluate this tool, six open source PHP applications were used. In PhpNuke, PhpMyAdmin and Gallery, 36 known vulnerabilities were recreated with 27 false positives. In Simple PHP Blog, Serendipity and Yapig, 15 unknown vulnerabilities were identified with 16 false positives. A restriction of Pixy is that it does not support 'OO' (object orientation).

![Table 6- Compare scanners](image)

<table>
<thead>
<tr>
<th></th>
<th>Path Traversal &amp; Local File Inclusion</th>
<th>Remote File Inclusion</th>
<th>Command Injection</th>
<th>Unrestricted File Upload</th>
<th>Open Redirect</th>
<th>CLRF injection</th>
<th>LDAP Injection</th>
<th>Xpath injection</th>
<th>SMTP/IMAP/EMAIL Injection</th>
<th>Server-Side Includes Injection</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6- Compare scanners

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>11</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>11</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>75.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>90.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>69.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>30.3</td>
</tr>
</tbody>
</table>
dependency graph which indicates the execution path and data flow obtained by the application to process the received request. Resp is the response returned by the application. \{query\} is a set of queries from the database generated by the application when processing a request. State, on the right, is the state to which the application transfers.

The implemented approach has three main components: dynamic analysis module that gathers the effects of application execution. Analyst that produces DDGs for gathered effects and the penetration test module that enters normal or malicious inputs into the application. For assessment, three applications in Python (Test application, Spyce and Trac) were used. In the presented results, false positive rate is zero.

In [26], a scanner is proposed to identify injection vulnerabilities. This system analyzes websites, aiming at automatically finding SQL and XSS injection vulnerabilities. The proposed system consists of two components: spider and scanner. The spider is used to navigate the site and find input points. The scanner initiates injection test and response analysis and consists of two parts: response analyst and author of rules. The system was run in VMware work station ACE with two hosts, one for the defense server and the other for the web server. The system was designed with PHP5 and MySQL and used the cURL module to execute attacks. Seven applications from National Vulnerability Database (NVD)\(^\text{11}\) were selected for assessment. Finally, the designed scanner was compared with some other scanners. It concluded that this system is effective, and vulnerability detection based on input points definitely can find vulnerabilities.

BLOCK is a black-box approach designed in [27] to identify state violation attacks based on the WebScrab tool. In this paper, the application is considered as a stateless system, and the application behavior model is obtained from the interaction between client and application, i.e. the relationships between web requests, responses and session variables. BLOCK has two key phases to identify state violation attacks: in the training phase, the desired behavior model is obtained by observing web request/response sequence, and the variable values corresponding to the session during its running without attacks. In the identification phase, the obtained model is used to evaluate each incoming web request and outgoing response, and any violation is identified. To assess this tool, the applications of Scarf, Simplecms, BloggIt, WackoPicko and OsCommerce were used. Results show that this method is effective in identifying state violation attacks and is incurred little overhead. Since this method is independent of the application code and can be used for a large number of web applications with different frameworks, it is of particular importance.

In [28], a mealy state machine is used to model web applications for management of requests that change the application mode. To detect a state change in this model, difference in the returned response in case of identical sent requests was considered. To implement this method, htmlUnit and the fuzzer of w3af tool were used. In the state graph the tool generates, nodes represent states and edges represent the requests sent to the application. Using the graph coloring problem, this article combined similar states. For evaluation, this product was studied with wget, w3af and skipfish scanners on the Gallery application, two versions of PhpBB, Scarf, vanilla forums, WackoPicko and three versions of wordPress. The evaluation results indicate that the designed scanner not only can run more codes than web applications, but also it can identify vulnerabilities not identified by other scanners. Due to the use of HtmlUnit, despite using the w3af fuzzer tool, this tool has a less false positive than other tools. Among the limitations of this tool, one can point to the lack of support of AJAX and applications that can be used publicly because users may influence the state-change detection algorithm.

In [29], an automatic black-box tool is presented to identify reflected XSS and stored XSS vulnerabilities in web applications. The tool uses user interactions to do the test more effectively. First, user interactions are recorded, then changes are made on these interactions for attacks and finally, this transaction is re-executed on the system. To evaluate the

\(^{11}\) A set of standards based on vulnerability management data owned by the US government which uses the Security Content Automation Protocol (SCAP).
performance, the tool was compared with Spider, Brup Spider, w3af and Acunetix on three applications of the Django framework from different aspects. Results show that the proposed method can identify more bugs than the listed commercial and open source tools.

In [30], a tool called MiMoSA is presented based on static analysis for PHP applications. The paper divides multi-module into two groups (data flow and workflow) and input states into two categories (server side and client side). The analysis intended to MiMoSA has two phases: intra-module which reviews each module of the application on its own, and inter-module which considers the whole application. The intra-module analysis aims to summarize each module of the application by defining preconditions, post-conditions and sinks. All links in each module are also extracted. This phase is dependent on the programming language. This information is then used for inter-module analysis to obtain the future state of the application workflow. In the inter-module analysis phase, results from intra-module analysis are put together in a single graph which models the future work flow of the whole application. Then a model survey technique is used to identify data flow vulnerabilities and future workflow defects. This phase is independent of the programming language used in the application development. To assess the tool, three applications of Aphpkb, BloggIt, MyEasyMarket, Scarf and SimpleCms, developed in the PHP language, were used. Evaluation results show that MiMoSA can identify all known vulnerabilities and also discover some new vulnerabilities. In evaluating the tool, only one false positive was seen. The number of states the tool considers is more than the number of actual states in the application code. This problem occurs for two main reasons: first, in MiMoSA, states may be produced correspond to paths that cannot be run in the application and second, the possibility of repetitive states in the tool with aligned but different conditions.

In [31], first the dynamic analysis and observing the application performance are used to understand the application's behavioral characteristics. Then the found characteristics are filtered to reduce false positive. The symbolic evaluation model is used on inputs to identify the application paths which violate these conditions. This paper focuses on logical vulnerabilities. The provided tool is Waler which is used for servlet-based application developed in Java. To assess the tool, 12 applications were used. According to this article, Waler is the first tool that can detect logical processes in applications automatically without human intervention.

There are many different tools for implementing each of the penetration test steps. In [32], for saving time, some of these tools were combined including theHarvester, Metagoofile, ZAP, NMAP, Nessus and Metasploit. The language used in this article is Python. The function of proposed method can be explained in three phases: information gathering, analysis of the information obtained and the use of this information to find possible vulnerabilities. First, the tools theHarvester, Metagoofile and NMAP are run as information gathering tools and Nessus and ZAP as two scanners, and information obtained by each tool is stored in a separate file. Then the results of ZAP and Nessus are analyzed and the resulting output file is used for implementing the attack phase on Metasploit.

The section addresses 10 studies that proposed a new method or tool for finding vulnerabilities in web applications. Each study is based on static or dynamic analysis, or a combination of the two. A summary of these characteristics is given in Table .

Table 1: papers proposed tools and methods summary

<table>
<thead>
<tr>
<th>Year</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>cross-site scripting detection</td>
</tr>
<tr>
<td>2008</td>
<td>identify errors in applications by path input detection</td>
</tr>
<tr>
<td>2008</td>
<td>Use TDM to identify internal vulnerabilities</td>
</tr>
<tr>
<td>2010</td>
<td>XSS and injection vulnerabilities detection</td>
</tr>
<tr>
<td>2011</td>
<td>identify state violation attacks based on the WebScrab tool</td>
</tr>
<tr>
<td>2012</td>
<td>Use state machine to model web application for detect more vulnerabilities</td>
</tr>
<tr>
<td>2008</td>
<td>reflected XSS &amp; stored</td>
</tr>
</tbody>
</table>
### XSS detection

<table>
<thead>
<tr>
<th>Year</th>
<th>Code analysis to models the future work flow of the application</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>✓</td>
</tr>
<tr>
<td>2010</td>
<td>✓</td>
</tr>
<tr>
<td>2013</td>
<td>✓</td>
</tr>
</tbody>
</table>

#### 3.3.3. Design of Test Environments

Vulnerable web applications are used to web vulnerability scanners evaluation. In this part we will have a review on some of this applications like Damn Vulnerable Web App (DVWA)\[33\], OWASP WebGoat\[34\], WackoPicko\[18\] and BodgeIt\[35\] that are used to training courses and numerous articles.

Damn Vulnerable Web App (DVWA) is a PHP/MySQL web application that is damn vulnerable. Its main goals are to be an aid for security professionals to test their skills and tools in a legal environment, help web developers better understand the processes of securing web applications and aid teachers/students to teach/learn web application security in a class room environment. Brute Force Login, Command Execution, CSRF, File Inclusion, SQL Injection, Upload Vulnerability and XSS are the vulnerabilities exist in DVWA\[33\].

Webgoat is a J2EE based web application designed by OWASP to introduce common security flaws in the web applications. The following categories are available in Webgoat: Access Control Flaws, AJAX Security, Authentication Flaws, Buffer Overflows, Code Quality, Concurrency, Cross-Site Scripting (XSS), Improper Error Handling, Injection Flaws, Denial of Service, Insecure Communication, Insecure Configuration, Insecure Storage, Malicious Execution, Parameter Tampering, Session Management Flaws, Web Services, and Admin Functions [34].

WackoPicko is a vulnerable website that designed by Adam Doupé and used in \[18\] for the first time. Vulnerabilities of this application are reflected XSS, stored XSS, SessionID vulnerability, stored SQL injection, reflected SQL injection, Directory Traversal, multi-step stored XSS, forceful browsing, Command-line Injection, File Inclusion, Parameter Manipulation, Reflected XSS Behind JavaScript, Logic Flaw, Reflected XSS Behind a Flash Form and Weak username/password.

BodgeIt is a web based vulnerable application designed by Simon Bennett with learning goals for pentesters. This web application consist of cross site scripting, sql injection, hidden (but unprotected) content, cross site request forgery, debug code, insecure object references, application logic vulnerabilities\[35\].

### Table 7: Test Bed Vulnerabilities

<table>
<thead>
<tr>
<th>Vulnerability</th>
<th>DVWA</th>
<th>WebGoat</th>
<th>WackoPicko</th>
<th>BodgeIt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brute Force</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>CSRF</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>File Inclusion</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SQL Injection</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Access Control Flaws</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XSS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Buffer Overflows</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Logic vulnerabilities</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>AJAX Security</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 4. Conclusion

The present paper reviewed studies in the field of penetration test, especially web penetration test. Manual penetration test is not effective in terms of time and money, so its automatic version is considered. For performing the automatic web penetration test, web scanners are used. They first crawl the target, then attack to the results of the previous phase and finally report vulnerabilities in the target. In this paper, we examined research in the field of web penetration test in three categories: articles that compared and analyzed available scanners, articles that proposed a new method or tool for penetration test and articles that proposed a test environment to test different tools. According to papers that analyzed various scanners, the Acunetix Web Vulnerability Scanner and IBM Rational AppScan scanners and the SQL injection and XSS vulnerabilities were considered more than others. We also reviewed 10 studies that proposed a new tool or method for penetration test, some of which were based on the dynamic analysis, some on the static analysis and some on a combination of the two. To
evaluate any method or tool in the field of penetration test, we require test environments. Four test environments were introduced in the final section.

The problems in existing scanners include the lack of support of attacks like stored sql and stored XSS that need to several steps to complete the attack, the lack of support of new technologies and vulnerabilities related to application logic flows. It is hoped that future work will consider these items.

References


[33] http://www.dvwa.co.uk/

[34] https://www.owasp.org/index.php/Category:OWASP_WebGoat_Project

[35] https://code.google.com/p/bodgeit/


[38] http://www.sectoolmarket.com/general-features-comparison-unified-list.html#Glossary


[41] Arefzadeh, A. (2012). Representing a penetration test plan for web-applications, based on RUP test plan(Master thesis, Department of Information Technology Engineering, Maleke-ashtar University of Technology)