

Natarajan Meghanathan
David C. Wyld (Eds)

Computer Science & Information Technology

7th International Conference on Computer Science, Engineering &
Applications (ICCSEA 2017)
September 23~24, 2017, Copenhagen, Denmark



AIRCC Publishing Corporation

Volume Editors

Natarajan Meghanathan,
Jackson State University, USA
E-mail: nmeghanathan@jsums.edu

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

ISSN: 2231 - 5403
ISBN: 978-1-921987-71-7
DOI : 10.5121/csit.2017.71101 - 10.5121/csit.2017.71114

This work is subject to copyright. All rights are reserved, whether whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the International Copyright Law and permission for use must always be obtained from Academy & Industry Research Collaboration Center. Violations are liable to prosecution under the International Copyright Law.

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The 7th International Conference on Computer Science, Engineering & Applications (ICCSEA 2017) was held in Copenhagen, Denmark, during September 23~24, 2017. The 9th International Conference on Wireless, Mobile Network & Applications (WiMoA-2017), The 6th International Conference on Signal, Image Processing and Pattern Recognition (SPPR-2017), The 9th International Conference on Grid Computing (GridCom-2017) and The 8th International Conference on Communications Security & Information Assurance (CSIA 2017) was collocated with The 7th International Conference on Computer Science, Engineering & Applications (ICCSEA 2017). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The ICCSEA-2017, WiMoA-2017, SPPR-2017, GridCom-2017, CSIA-2017 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, ICCSEA-2017, WiMoA-2017, SPPR-2017, GridCom-2017, CSIA-2017 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the ICCSEA-2017, WiMoA-2017, SPPR-2017, GridCom-2017, CSIA-2017.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Natarajan Meghanathan
David C. Wyld

Organization

General Chair

David C. Wyld
Jan Zizka

Southeastern Louisiana University, USA
Mendel University in Brno, Czech Republic

Organizing Committee

Brajesh Kumar Kaushik
Dhinaharan Nagamalai
Natarajan Meghanathan
Salah M. Saleh AL-MAJEED
Jae Wnag Lee

Indian Institute of Technology - Roorkee, India
Wireilla Net Solutions PTY LTD, Australia
Jackson State University, USA
University of Essex, United Kingdom
Hannam University, South Korea

Program Committee Members

Abbas Akkasi
Adnan Albar
Akhil Garg
Ali Hussein Mohammed
Alireza Afshari
Amel Boufrioua
Amir baharvandi
Barbaros Preveze
Chin-Chih Chang
Dac-Nhuong Le
Daniel D. Dasig
Ehsan Saradar Torshizi
Farhad Soleimanian
Gammoudi Mohamed Mohsen
Guo Yue
Hacene Belhadeif
Hamdi hassen
Hossein Jadidoleslamy
Irving V Paputungan
Isa Maleki
Jaesoo Yoo
Jamaiah Yahaya
Jamshid Aghaei
Javed Mohammed
Jeremy Briffaut
Jungpil Shin
Jyotsna Kumar Mandal
Ka Ching Chan
Laudson Souza

Islamic Azad University of Bandar, Iran
King Abdulaziz University, Saudi Arabia
Nanyang Technological University (NTU), Singapore
Alexandria University, Egypt
Islamic Azad University, Iran
University Constantine 1, Algeria
IEEE Transaction on Power Systems, Iran
Çankaya University, Turkey
Chung Hua University, Taiwan
Haiphong University, Vietnam
Jr., Jose Rizal University, Philippines
Urmia University, Iran
Hacettepe University, Turkey
University of Manouba, Tunisia
Ningbo University of Technology, China
University of Constantine 2, Algeria
Miracl laboratory, Tunisia
MUT University, Iran
Universitas Islam, Indonesia
Islamic Azad University, Iran
Chungbuk National University, Korea
The National University of Malaysia (UKM), Malaysia
Shiraz University of Technology, Iran
NewYork Institute of Technology, USA
INSA CVL/LIFO, France
University of AIZU, Japan
University of Kalyani, India
La Trobe University, Australia
Integrated Faculties of Patos (FIP) - Brazil

| | |
|-------------------------|---|
| Laura Felice | Universidad Nacional del Centro. Tandil. Argentina |
| Mohammad Jafarabad | Qom University, Iran |
| Muhammad Baqer Mollah | United International University, Bangladesh |
| Murat TOPALOGLU | Trakya University, Turkey |
| Nisheeth Joshi | Banasthali University, India |
| Noudjoud KAHYA | Badji Mokhtar University, Algeria |
| Ramgopal Kashyap | Sagar Institute of Science and Technology, India |
| Rkia Aouinatou | Mohammed V University, Morocco |
| Rommel Anacan | Technological Institute of the Philippines, Philippines |
| Saif Al-Alak | Babylon University, Iraq |
| Sangita Zope-Chaudhari | A. C. Patil College of Engineering, India |
| Sergio Takeo Kofuji | University of Sao Paulo, Brazil |
| Seyyed Reza Khaze | Islamic Azad University, Iran |
| Shamala Subramaniam | Universiti Putra Malaysia, Malaysia |
| Shamim H Ripon | East West University, Bangladesh |
| Shengwei Yu | InnoGrit Tech Inc., USA |
| Sukanyathara J | APJ Abdul Kalam Technological University, India |
| Taher M. Ali | Gulf University for Science & Technology, Kuwait |
| Varun Vohra | University at Buffalo, USA |
| Wenzhao Zhang | North Carolina State University, USA |
| Yassine MALEH | Hassan 1st University, Morocco |
| Yingchi Mao | Hohai University, China |
| Zenon Chaczko | University of Technology, Australia |
| Zid youssef | NEST, Tunisia |
| Zuriati Ahmad Zukarnian | University Putra Malaysia, Malaysia |

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)



Artificial Intelligence Community (AIC)



Organized By



Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

7th International Conference on Computer Science, Engineering & Applications (ICCSEA 2017)

| | |
|---|-----------|
| CINNAMONS : A Computation Model Underlying Control Network Programming..... | 01 - 20 |
| <i>Kostadin Kratchanov</i> | |
| Representation with Word Clouds at the PhD UNS Digital Library..... | 21 - 33 |
| <i>Georgia Kapitsaki and Dragan Ivanović</i> | |
| Usability Testing of Fitness Mobile Application : Methodology and Quantitative Results..... | 97 - 114 |
| <i>Ryan Alturki and Valerie Gay</i> | |
| Recognition the Droplets in Gray Scale Images of Dropwise Condensation on Pillared Surfaces..... | 115 - 126 |
| <i>Helene Martin, Solmaz Boroomandi Barati, Jean-Charles Pinoli, Stephane Valette and Yann Gavet</i> | |
| Predicting Software Launch Readiness in a Complex Product..... | 127 - 138 |
| <i>Abhinav Sharma</i> | |
| An Innovative Social Mobile Platform to Support Real Time Communication in Peer Tutoring..... | 161 - 170 |
| <i>Meghan Wang and Yu Sun</i> | |
| An Intelligent Self Adaptive System to Automate the Sprinkler Control..... | 171 - 176 |
| <i>Jiahao Li, Yu Sun and Fangyan Zhang</i> | |

9th International Conference on Wireless, Mobile Network & Applications (WiMoA-2017)

| | |
|---|---------|
| Attack Analysis in Vehicular Ad Hoc Networks..... | 35 - 46 |
| <i>Ömer Mintemur and Sevil Sen</i> | |
| Analysis of Wormhole Attack Confirmation System During Email Dumping Attack..... | 47 - 56 |
| <i>Divya Sai Keerthi T and Pallapa Venkataram</i> | |

Lightweight Key Management Scheme for Hierarchical Wireless Sensor Networks..... 139 - 147
Mohammed A. Al-taha and Ra'ad A. Muhajjar

6th International Conference on Signal, Image Processing and Pattern Recognition (SPPR-2017)

Extraction and Refinement of Fingerprint Orientation Field..... 57 - 68
Pierluigi Maponi, Riccardo Piergallini and Filippo Santarelli

9th International Conference on Grid Computing (GridCom-2017)

Adapted Bin Packing Algorithm for Virtuals Machines Placement into Datacenters..... 69 - 80
Fréjus A. R. Gbaguidi, Selma Boumerdassi and Eugène C. Ezin

8th International Conference on Communications Security & Information Assurance (CSIA 2017)

Cryptographic Strength Estimation Using Spurious Keys with Consideration to Information Content in the Message..... 81 - 95
Mekala Rama Rao, L Pratap Reddy, BHVS Narayana Murthy and Maruti Sairam Annaluru

Block Chain Based Data Logging and Integrity Management System for Cloud Forensics..... 127 - 137
Jun Hak Park, Jun Young Park and Eui Nam Huh

CINNAMONS: A COMPUTATION MODEL UNDERLYING CONTROL NETWORK PROGRAMMING

Kostadin Kratchanov

Department of Software Engineering, Yaşar University, Izmir, Turkey

ABSTRACT

We give the easily recognizable name “cinnamon” and “cinnamon programming” to a new computation model intended to form a theoretical foundation for Control Network Programming (CNP). CNP has established itself as a programming paradigm combining declarative and imperative features, built-in search engine, powerful tools for search control that allow easy, intuitive, visual development of heuristic, nondeterministic, and randomized solutions. We define rigorously the syntax and semantics of the new model of computation, at the same time trying to keep clear the intuition behind and to include enough examples. The purposely simplified theoretical model is then compared to both while-programs (thus demonstrating its Turing completeness), and the “real” CNP. Finally, future research possibilities are mentioned that would eventually extend the cinnamon programming into the directions of non determinism, randomness and fuzziness.

KEYWORDS

Control network programming, CNP, Programming languages, Programming paradigms, Computation models, While programs, Theoretical computer science, Recursive automata, Non determinism, Semantics

1. INTRODUCTION

This paper introduces a new computation model called **Core Control Network Programming**, or **Core CNP**. We introduce it to serve as a theoretical basis for Control Network Programming [1-4]. Actually, to make the new computation model easier to distinguish (and influenced by the popularity of Pokémon), we will slightly twist the name “Core CNP” into “**Cinnamon programming**”. As it is intended for formal mathematical study, it is a strictly minimal version of CNP - it has the same structure, but with only the most basic and fundamental features and primitives necessary. Sections 2 and 3 are devoted to the syntax and semantics of the new computation model, respectively. In Section 4 we show the Turing-completeness of cinnamon programming. Section 5 includes a concise description of more advanced features that distinguish the “real” CNP from the core CNP while Section 6 sketches some ideas for future study.

In addition to developing theoretical foundations for CNP, an equally important aim of the presentation is to make the syntax and semantics of Control Network (CN) programs unambiguous and clear for the programmers employing CNP.

1.1. Background and Perspective

As Jones explicitly maintains in his Preface to [5], the two major computer science fields – Theory of Computation (which splits into Computability and Computational complexity), and Programming languages (including Syntax and Semantics) – have much to offer each other. This overlap of concepts, approaches, and results can be described as a trend observed in the recent decades, some examples being [6-8]. We would like to follow this trend here and use the terminology from both areas.

The purpose of a computation is to solve a **problem**. Clearly, we must first make precise what is meant by a problem (also called **computational task**). Currently, there are four well-established main types of computational tasks considered in the computer literature: decision problems, function problems, search problems, and optimization problems.

There are numerous approaches to the definition of **computation** and, respectively, to answering the question what is computable. Examples of some of the most famous approaches are: recursive functions (originating from the notion of functions in mathematics and based on the intuitive idea of what operations can be reasonably considered to be easily computable), lambda calculus (underlying functional programming), Turing machines (a maximally simplified version of a computer), register machines and RAM machines (underlying machine languages and assembly languages), GOTO-programs and WHILE-programs (theoretical foundation for higher-level imperative programming languages), first-order logic (underlying logic programming). A really striking result and one of the most important achievements of computer science is the fact that all these very different approaches ultimately lead to the same classification of decision problems, respectively function problems, with respect to their computability. Computability (solvability) turns out to be a natural intrinsic property of problems independent of the formalism used in the corresponding approach. These formalisms are referred to as models of computation. A **computation model** is called Turing-complete if it is equivalent in its computational power to a Turing machine (and, correspondingly, to any other of these universal models).

Under each of these approaches, a model of computation is defined as a formal mathematical object, followed by the concept of computation for that model and the notion of what a given specific model computes. Assume that Turing machines are our computation model. Turing machines – acceptors are used for solving decision problems, while Turing machines – transducers are used for solving function problems. A Turing machine – acceptor solves a decision problem by accepting or rejecting the element. A particular Turing machine – transducer calculates a partial function by producing an output element for an input element.

Note that, ultimately, every Turing machine (including its program), or WHILE-program, etc. has a unique description in some properly defined language. Therefore, we can talk about the language of Turing machines, the language of WHILE-programs, and so on. A particular Turing machine is an element (a well-formed sentence) of the language of Turing machines.

The description of a language is split into **syntax** (form) and **semantics** (meaning). Semantics reveals the meaning of syntactically valid strings in a language. There is a wide range of semantics proposed for programming languages (see, e.g., [4]). Two of the major approaches are outlined next. **Operational semantics** describes how a computation is performed internally in the corresponding computation model (often also called an abstract machine) – that is, how the program is interpreted (executed) in the computation model. In contrast, in **denotational semantics** we give meaning to the program by specifying the external behavior achieved by the computation, e.g., the partial function computed.

The above means that introducing a formal model of computation as a mathematical object (e.g., Turing machines) and defining how the computation in its terminology is performed (how a Turing machine operates and what it computes) is actually equivalent to introducing a language (the language of Turing machines) and specifying its syntax and semantics.

1.2. Cinnamons

This paper introduces a new computation model called **Core Control Network Programming**, or **Core CNP**.

What in traditional programming corresponds to a program, is called here a **cinnamon** (the formal definition of a cinnamon is given in the next section). A traditional program is a string in the corresponding programming language. In contrast, a cinnamon is not simply a string but rather a ‘control network’ – a set of graphs. Cinnamon programming is a type of graphical (visual) programming. We believe graphical programming is clearer and more natural for human programmers [1,4], and this results in a faster process of developing a solution to the problem in hand. Of course, at a lower level the CN is transformed into a string in an appropriate language which is manipulated by the CNP compiler. However, in this presentation we prefer to stay at the higher and more abstract level.

It is customary in mathematics to discuss objects without specifying their representation (talking about sets of objects, functions between sets of objects, etc.). In computation theory the representation of objects often plays a central role [10]. However, for clarity and simplicity, we will still define, as much as we can, the notions and concepts we need using sets without regard for the representation of their elements. Actually, computational tasks refer to objects that are represented in some canonical way. The two such main and best studied alternative representations are strings of symbols and natural numbers. In our presentation, when that must be made explicit we have chosen to work with natural numbers, and correspondingly functions on natural numbers.


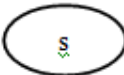
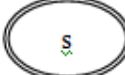


2. SYNTAX

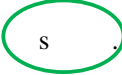
The syntax of the “language” of cinnamons is defined below.

1.2. Definitions

Let $SV = \{FINISH, RETURN\}$ be a set of two distinct elements called **system vertices**(**system nodes, system states**), $FINISH \neq RETURN$. A **graph** is an ordered 6-tuple $G = \langle S, A, source, target, L, label \rangle$ consisting of: a nonempty finite set, S of elements called **ordinary states** (**ordinary nodes, ordinary vertices**), $S \cap SV = \emptyset$; a finite set, A of **arrows**; two functions $source: A \rightarrow S$ and $target: A \rightarrow (S \cup SV)$ mapping an arrow $a \in A$ into its **sources** $\in S$ and **targets** $\in (S \cup SV)$ respectively; a set, L of **labels**; and a function $label: A \rightarrow L \cup \{\lambda\}$ assigning a label $label(a) \in L \cup \{\lambda\}$ to each arrow $a \in A$. A **state** is either an ordinary state, or a system state. For a given $s \in S$, we will denote by $out(s)$ the set of all arrows with source s (called also arrows outgoing from s). An **initialized graph** is an ordered pair $\langle G, s_o \rangle$ where G is a graph and s_o is a selected state, $s_o \in S$ called the **initial state**. Note that $FINISH$ and $RETURN$ have no outgoing arrows, and that these system states cannot be initial. We will call a graph, G an **ordered graph** if the set $out(s)$ for any given vertex $s \in S$ is linearly ordered. An ordered graph which is initialized, is called an **ordered initialized graph**.

We will represent initialized graphs graphically, using the following graphical symbols:

| | | | | | |
|-------------------------|---|---|---|---|---|
| Graphical symbol |  |  |  |  |  |
| Element | Arrow with label l | (Ordinary) vertex | Initial vertex | <i>FINISH</i> vertex | <i>RETURN</i> vertex |

In the *SpiderCNPIDE*, the graphical symbol for an initial node is an oval in green: .

SpiderCNP [4,11,12] is an integrated development environment for developing and running CN programs. In particular, CNs can be created, tested and edited using a built-in graphic editor. All illustrations of CNs in this paper are screenshots from this IDE. *SpiderCNP* was developed by T. Golemanov.

A **subnet** is an ordered pair $\langle G, P \rangle$ where G is an ordered initialized graph, and P is a list (possibly empty) with elements called **formal subnet parameters**.

Let $Vars$ be a countable set. Its elements will be called **variables**. We will denote the elements of Var as x_0, x_1, x_2, \dots . Of course, $Vars$ could be the set N of natural numbers itself.

A **cinnamon** (a substitution for Simple CN) is comprised of a nonempty finite set of subnets one of which is identified as the **main subnet**. Formally, a cinnamon, Σ is an ordered pair $\langle N, m \rangle$ where N is a set of subnets satisfying conditions 1) - 2) below, and $m \in N$ is its main subnet.

This set of subnets must satisfy the following two properties:

- 1) The sets of vertices of the (graphs of the) subnets are mutually nonintersecting.
- 2) All subnets share the same label set, L . This label set consists of two types of elements: **primitives**, and **invocations** (the latter also called **subnet calls**) described below.

Each primitive in the cinnamon may be from one of the following four types:

- a) *clear*(x) (also denoted $x:=0$)
- b) *copy*(x,y) (also denoted $y:=x$)
- c) *inc*(x) (also denoted $x:=x+1$)
- d) *if nonEq*(x,y) (also denoted *if* $x \neq y$)

Primitives *clear*, *inc*, and *copy* are called **elementary action primitives**. Primitive *if nonEq* is called the **elementary test primitive**.

Above, x and y are variables. Note that Σ consists of a finite set of subnets, each of which has a finite set of labeled arrows. Therefore, the number of variables used in it is also finite. We will denote by $Vars^\Sigma$ the set of all variables used in the primitives of the cinnamon Σ . If η is a subnet of a cinnamon Σ , then $Vars^\eta$ denotes the set of variables of η . The sets of variables of two distinct subnets are nonintersecting. Clearly, $Vars^\Sigma$ is the union of $Vars^\eta$ for all subnets η of Σ , $Vars^\Sigma$ is finite, and $Vars^\Sigma \subset Vars$.

An invocation has the form:

- 3) *CALL* $\eta(a_0, a_1, \dots, a_{n-1})$ where η is a subnet of the cinnamon Σ , and $a_0 - a_{n-1}$ are called **actual subnet parameters**. Each a_i is either a variable from $Vars^\Sigma$, or a constant (a natural number).

The list of actual subnet parameters may be empty, in which case we simply write *CALL* η . The number of actual parameters in the subnet call equals the number of formal parameters in the subnet definition.

Both direct and indirect recursion are allowed. For example, it is possible that the invocation $CALL(\eta)$ is located in subnet η , that is, subnet η is called from itself.

According to our definition above each arrow has a single label which is an elementary primitive or a subnet call. This assumption is imposed for simplicity. In practice, it is more convenient to allow an arrow's label to be a finite, possibly empty, string of primitives and/or invocations. If

using the new notation, $S \xrightarrow{p_1, p_2, \dots, p_n} S'$ actually denotes $S \xrightarrow{p_1} S_1 \xrightarrow{p_2} \dots \xrightarrow{p_n} S'$

and $S \xrightarrow{\lambda} S'$ (empty arrow) denotes $S \xrightarrow{\lambda} S'$

An example of a cinnamon is shown in Figure 1.

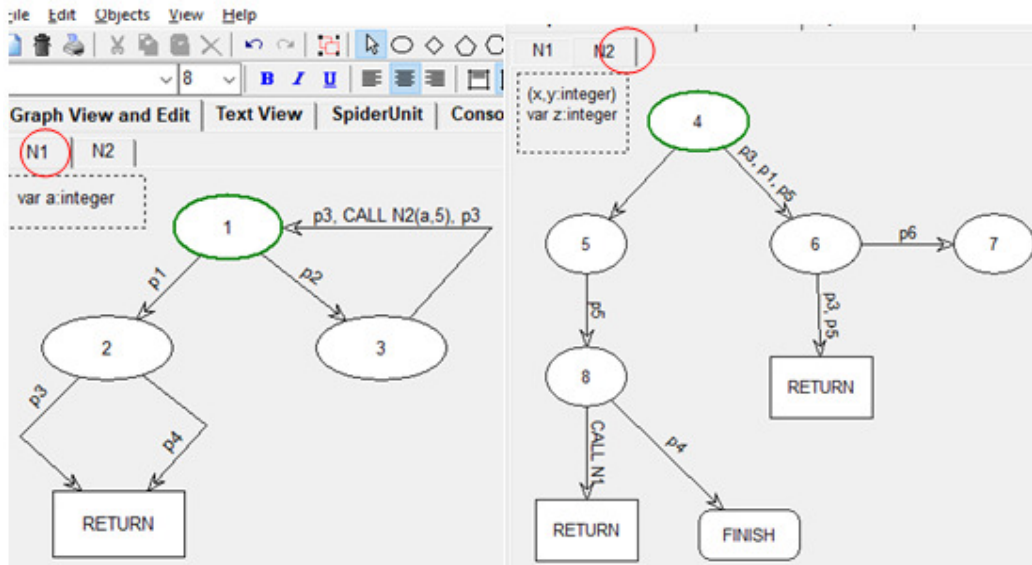


Figure 1 Cinnamon example 1

Here, the cinnamon consists of two subnets, $N1$ and $N2$. $N1$ is the main subnet. It includes the ordinary states $\{1, 2, 3\}$, as well as a $RETURN$ system state. The initial state is 1 . Primitives used in $N1$ are $\{p1, p2, p3, p4\}$; the invocation $CALL N2(a,5)$ is also used. In the invocation, a is an actual subnet parameter; the second parameter in this call is the constant 5 . Subnet $N2$ uses the primitives $\{p1, p3, p4, p5, p6\}$, and the invocation $CALL N1$. Its set of states is $\{4, 5, 6, 7, 8\}$. $N2$ also includes two $RETURN$ states and one $FINISH$ state. Each one of the primitives $p1 - p6$ must be of the forms discussed above, i.e., either an elementary action primitive or elementary test primitive. Subnet $N1$ has no formal subnet parameters while subnet $N2$ has two formal subnet parameters, x and y . This is an example of indirect recursion: subnet $N2$ is called from within subnet $N1$, and $N1$ is called from within $N2$. Clearly, invoking the main subnet from the same or other subnet is not a hindrance.

As emphasized, the arrows going out of a given node, are linearly ordered. For convenience we accept that on the graphical representations the order of the arrows is from the left to the right and up - down. For example, the arrow $1 \rightarrow 2$ in $N1$ is before arrow $1 \rightarrow 3$. If using this default rule on the graphical representation is difficult, we may also show the order explicitly, as in Fig. 2. Here, arrow $1 \rightarrow 3$ precedes arrow $1 \rightarrow 2$.

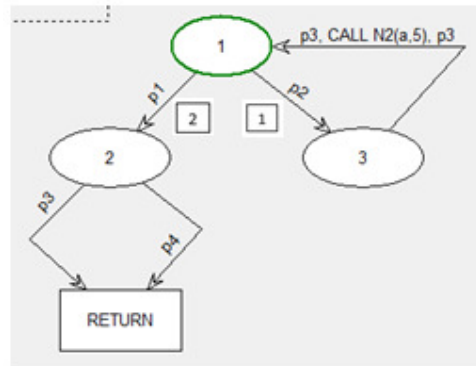


Figure 2 Indicating order of arrows

In our definition of a cinnamon above we have chosen a minimal set of allowed elementary primitives consisting of the specified three elementary action primitives ($x:=0$, $y:=x$, and $x:=x+1$) and one test primitive ($if\ x\neq y$). As we will see further, this minimal set is functionally complete in the sense that these primitives are enough for implementing any computation. It would have been possible to choose alternative sets, e.g., action primitives $\{x:=0, y:=x+1, y:=x\div 1\}$. As far as the test primitive is concerned, its condition could have been replaced by any one of conditions $=, <, >, \leq, \geq$ (similarly to [6,13] in the case of while-programs). For our considerations, however, our choice of elementary primitives proves to be the most convenient.

It is also possible to choose a different set of elementary primitives if working with a based-on-strings representation of data. Let A be a fixed alphabet with at least two different symbols. We could, for example, choose the following set of elementary primitives: $clear(l)$ (creating an empty string $l=\lambda$), $sep(l,h,t)$ (separating a string l into its head h and tail t), $cons(h,t,l)$ (constructing a new string l with head h and tail $t =$ adding a new symbol to the beginning of a string), $ifEq(a,a1)$ (comparing symbols), $ifEmpty(l)$ ($if\ l = \lambda$).

It is worth mentioning that the concept of a cinnamon is an elaborated, filled with flesh version of an introduced much earlier but not so widely used skeletal notion – the notion of a recursive automaton / recursive finite-state automaton / recursive transition network / recursive control graph [14-23]. It is a computation model equivalent in computational power to nondeterministic push-down automata. Basically, the difference between an automaton and a recursive automaton is that the latter is a set of nondeterministic automata in which transitions are labeled either by an element of the input alphabet (a primitive in the case of cinnamons) or by the name of a nondeterministic automaton from the set (a subnet in the case of cinnamons). Cinnamons are a more complicated concept than recursive automata – for example, primitives and subnets have parameters, the set of outgoing arrows is ordered, the behavior in the basic definition is deterministic, there are concepts like failure and backward execution (the latter differences are of semantic nature).

As is customary in computation theory literature, we can allow for convenience the usage of macro primitives. A **macro definition** is a sequence of primitives. For example, the sequence $\langle clear(z), nonEq(x,z) \rangle$ can be considered as a macro definition for the macro statement $nonZero(x)$ which would be a new test primitive. Such an abbreviation will be called a **macro statement** (plural: **macros**).

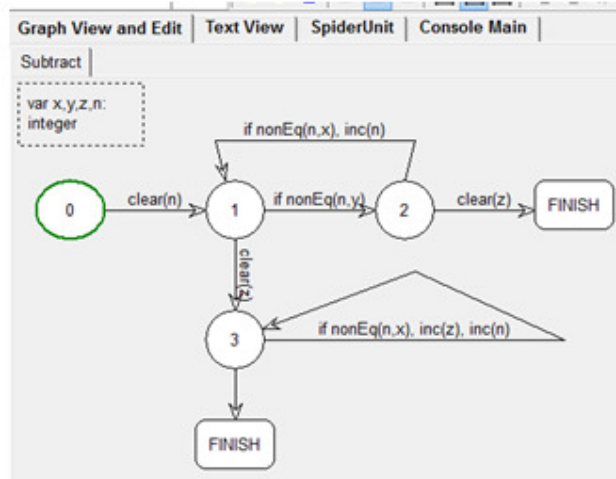
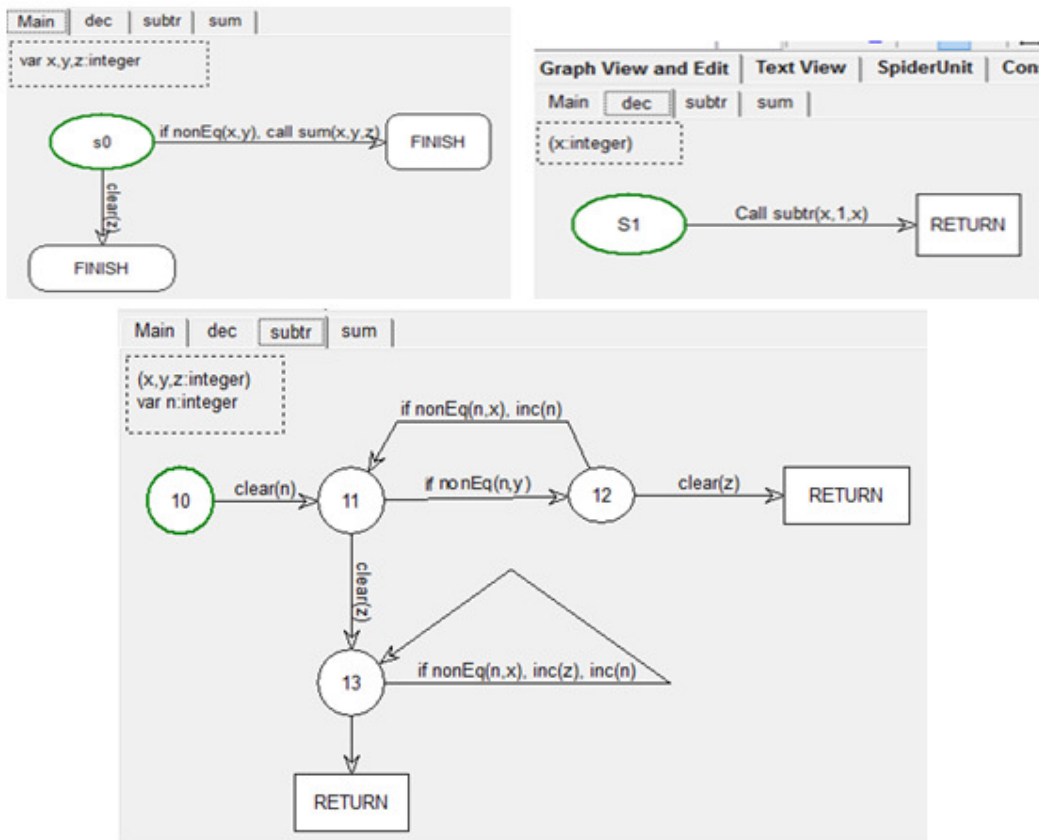


Figure 3 Cinnamon for subtraction

2.2 Examples

A second example of a cinnamon is shown in Fig. 3. It consists of a single (main) subnet *Subtract*. After understanding the semantics (computation) of cinnamons, one can convince himself that this cinnamon computes the function subtraction of natural numbers; more precisely,

$$z = \begin{cases} x - y, & \text{if } x \geq y; \\ 0, & \text{if } x < y. \end{cases}$$



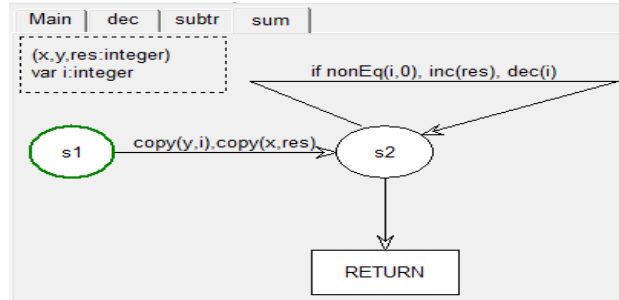


Figure 4 A larger example

Macros are one method for separating parts of a cinnamon into a named module of some sort and re-use them. Another, more powerful method is using subnets. In the following example (Figure 4), we use a modified version of the previous example – the cinnamon for subtraction, but now it is not a complete cinnamon but rather a subnet within a cinnamon. The cinnamon has four subnets. One is subnet *subtr* already discussed. Subnet *dec* ‘defines’ a new ‘operation’ – decrement, using the subnet for subtraction. Using *dec*, subnet *sum* ‘defines’ another new ‘operation’ – summation. The overall control structure is determined by the main subnet. This example actually implements the function

$$\text{if } x \neq y \text{ then } z := x + y \text{ else } z := 0$$

Note that ordering of outgoing arrows from a certain node plays an important role. For example, according to the default ordering of arrows outgoing from node *s0* in subnet *main* arrow $s0 \rightarrow \text{FINISH}$ labelled *if nonEq(x,y), call sum(x,y,z)* precedes the arrow labeled *clear(z)*, and the loop $s2 \rightarrow s2$ in subnet *sum* precedes arrow $s2 \rightarrow \text{RETURN}$.

Our next example (Fig. 5) illustrates a cinnamon based on the definition with data representation using strings instead of natural numbers. Here, *copy(t,l)* is a macro statement with macro definition $\langle \text{sep}(l,h,t, \text{cons}(h,t,l)) \mid (l := t) \rangle$.

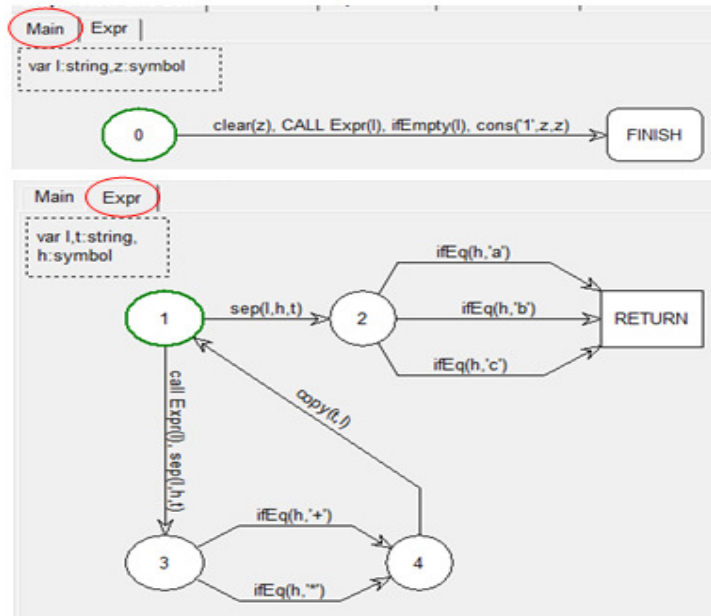


Figure 5 Recognizer of arithmetic expressions

Assuming that the alphabet A includes symbols $a, b, c, +$, and $*$, this cinnamon plays the role of a recognizer for arithmetic expressions defined by the (ambiguous) grammar

$$E \rightarrow E + E / E * E / a / b / c$$

Equivalently, this grammar can be defined by the syntax diagram in Fig. 6. The structure of the graph of the cinnamon duplicates that of the syntax diagram. The left recursion has been avoided.

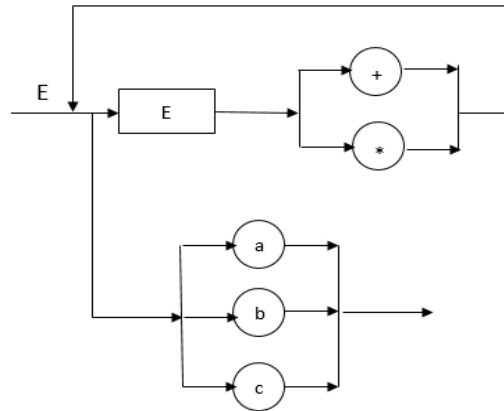


Figure 6 Syntax diagram of the grammar of arithmetic expressions

Note that the list of elementary primitives in the definition of a cinnamon includes no ‘control’ primitives. This contrasts with the other well-known computation models such as while-programs, recursive functions, etc. where one can find ‘control’ components such as while-loops, conditional statements, functions of primitive recursion and unbounded minimization, etc. The primitives in our definition are elementary tests or actions, and the control is embodied in the structure of the subnets. In the ‘real’ CNP, however, primitives are user-defined and can be arbitrary procedures defined in the underlying language – they can include loops, etc. In the triple general computational model terminology [22], variables and primitives define the operational unit, while the CN defines the control unit. As emphasized in [4], “Primitives + Control Network = Control Network Program”. In cinnamon programming primitives are built-in. In real CNP primitives must be defined and therefore a CNP project in *SpiderCNP* or the *Bouquetcloud* development environment [4] involves two major files – *SpiderUnit* and *SpiderNet*; in cinnamon programming *SpiderUnit* is not needed.

Above, a cinnamon was defined abstractly using purely mathematical notions (such as sets and graphs) and notations. Alternatively, we could’ve had defined a formal language, or a programming language in which a cinnamon is specified. For our purposes in this paper the more abstract approach is preferable as it presents the concepts in a much more clear and neat manner. In reality, for developing and running CN programs one usually uses an integrated development environment such as *SpiderCNP* or *Bouquet* [4]. In *SpiderUnit*, two equivalent representations of the CN exist – a graphical representation and a corresponding textual representation. Typically, the ‘programmer’ creates and modifies the CN using its graphical representation manipulated by the built-in graphical editor, while the textual representation of the same CN is used in the files with which the CNP compiler and other software modules work.

3. SEMANTICS

In the previous section we presented the syntax of the ‘language’ of cinnamon programming, in other words, what syntactically a cinnamon is. As already discussed this is technically not a language as a cinnamon, unlike a program, is not a string but rather a set of graphs.

Now we turn our attention to the semantics of a cinnamon – in other words, what a cinnamon ‘computes’. This specification of the computation process can be considered as an operational semantics. One can also consider a partial function which is defined by this computation thus saying that a cinnamon as a syntactic construct specifies this function – and that can be considered as a denotational semantics.

3.1 Informally on the computation in a cinnamon

We will describe first the ‘computation’ performed in a cinnamon, and then will address the semantics of the model more rigorously. Our exemplary cinnamon (which is a simplified version of the ‘technical example’ from [22]) is shown below.

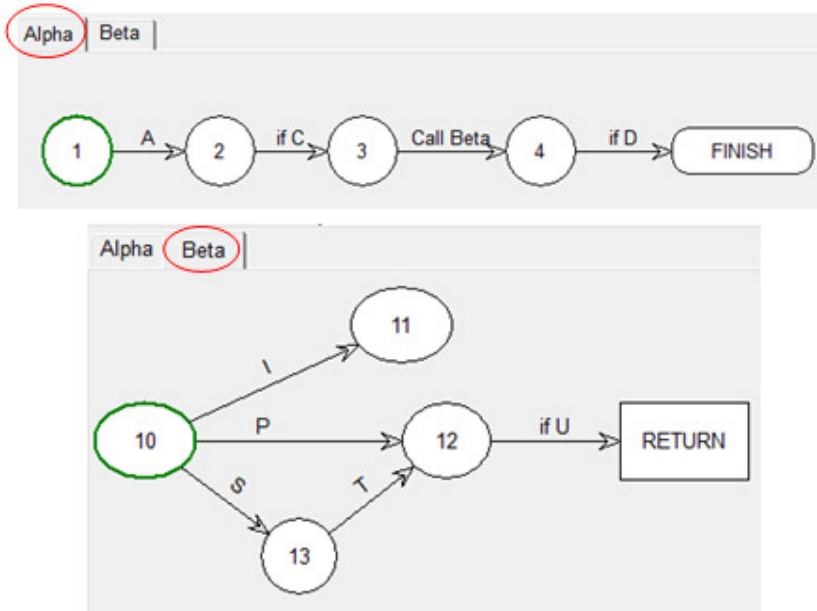


Figure 7 Technical example

The computation starts from the initial state of the main subnet by ‘forward’ traversal along the arrows, and executing each primitive on the way. An elementary test primitive can be executed successfully, or unsuccessfully. After unsuccessful execution, the direction of the traversal is switched to ‘backwards’. How a primitive is executed forwards and backwards is given in its definition. The traversal strategy is an extended version of backtracking. When the control is returned to a state, the next not attempted yet outgoing arrow is tried. If no non-attempted outgoing arrows exist, the control goes backwards through the arrow along which the state was reached. A very interesting point of cinnamon programming is that the backward traversal can enter backwards a subnet, and continue backwards along the arrow used before. This is not possible in traditional programming because when completing a procedure all information from inside the procedure is lost. In contrast to traditional procedure/function call in programming languages, in cinnamon programming the data are not restored from a data stack but are restored by backwards execution. If the traversal reaches a *FINISH* node then the computation finishes successfully. If the traversal gets stuck in the initial node of the main subnet then the computation finishes unsuccessfully – no solution has been found.

One possible traversal is shown in Fig. 8. The execution starts from the initial state, *I* of the main subnet, *Alpha*. Action primitive *A* is executed forwards, then test primitive *if C* is executed successfully. Next invocation *Call Beta* is executed, after which the control is in the initial state, *I0* of subnet *Beta*. Primitive *I* is now executed, and control moves to state *I1*. However, state *I1*

has no outgoing arrows, and mode of traversal changes to ‘Backwards’. Primitive I is now executed backwards and control is back in state 10 . The next in order not-attempted outgoing arrow is the one with primitive P which is now executed. Then test primitive *if U* is successfully executed, *RETURN* state is reached, and control jumps back to subnet *Alpha*, more specifically to state 4 . Test primitive *if D* follows, but its execution is unsuccessful. Therefore, backwards execution is triggered. Primitive *if D* is executed backwards, and control returns to state 4 . There are no other outgoing arrows from 4 . Therefore, subnet *Beta* is entered backwards. The system remembers that the last executed arrow before returning from subnet *Beta* was the one with label *if U*. Now, if U is executed backwards. State 12 has no remaining not-attempted arrows, so control continues moving backwards and primitive P is executed backwards. Now the control is in state 10 . Outgoing arrow S has not been attempted, so the mode changes back to ‘Forward’ and primitive S , and then T and *if U* are executed forwards. Now, return from subnet *Beta* is performed and control reaches state 4 . Assume that test primitive *if D* is now successful. Thus, the control reaches state *FINISH*, and the computation completes successfully.

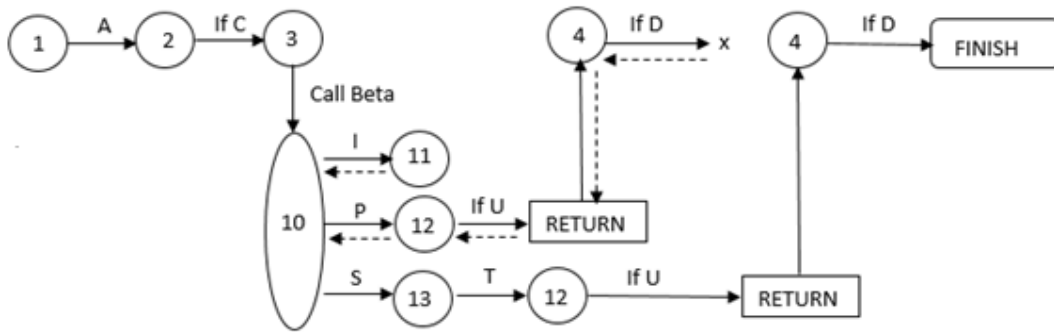


Figure 8 One possible execution in the technical example

Now we can proceed to a formal definition of the cinnamon semantics.

3.2 Formal semantics

Let $Vars$ be the countable set of variables defined earlier in the Section on Syntax. Let Σ be a given cinnamon, and $Vars^\Sigma \subset Vars$ be its finite set of variables. As a matter of fact, we never use values of $Vars$ outside of $Vars^\Sigma$ but the definitions remain simpler for Var . We will also need two special ‘system variables’ *FORW* and *FAILURE*. We assume that $SV = \{FORW, FAILURE\}$, $SV \cap Vars = \emptyset$, and $Varss = SV \cup Vars$.

An **environment**, ε is a partial function $Varss \rightarrow N$. Terms with similar meaning used in literature are state of computation, store, single-assignment store. If for $v \in Varss$ the partial function σ is defined then the natural number $\varepsilon(v)$ is called the **value** of variable v . If $\varepsilon(v)$ is not defined, we will also say that the value of variable v is not defined. The values $\varepsilon(FORW)$ and $\varepsilon(FAILURE)$ can only be 0 or 1, therefore the two system variables will more often be called **system flags**. The set of all environments is denoted Env .

Informally, $Vars^\Sigma$ imitates the data store (operational unit) in our model. We also need a control unit which in cinnamon programming is much more complex than in most other computation models as cinnamon programming is intended for declarative solution of problems of nondeterministic nature. Following our general strategy in this paper, we will employ mathematical definitions rather than explicitly specifying the abstract machine or use programming style terminology. However, in order to simplify the presentation, we will still use popular notions such as a stack of a particular type of elements for example, instead of describing it in formal mathematical language as a function $N \rightarrow Elements$.

We will define the operational semantics of cinnamons by presenting formally an interpreter. The algorithm of this interpreter is specified by its UML activity diagram shown in Figure 9.

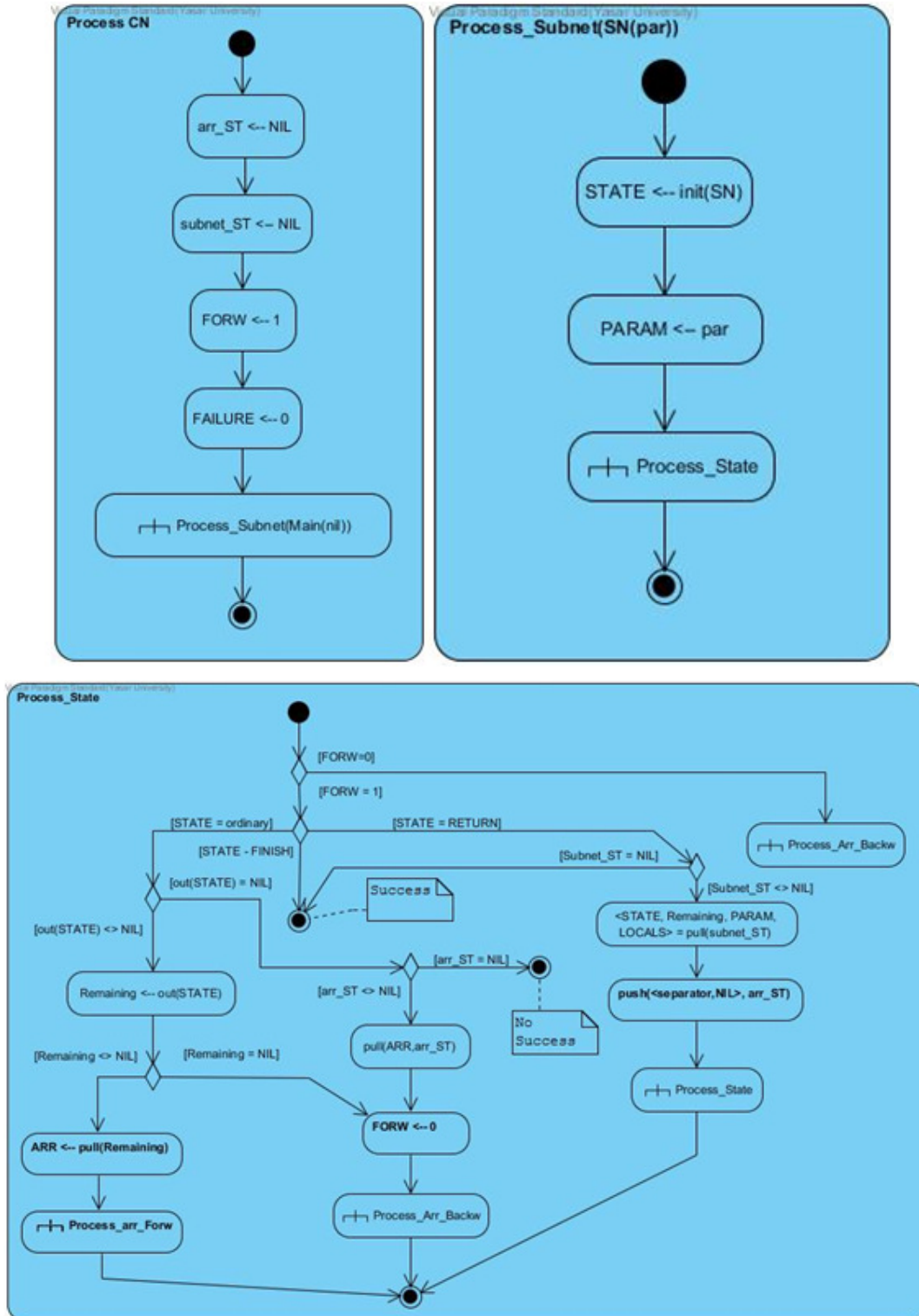


Figure 9a Activity diagram of the interpreter I

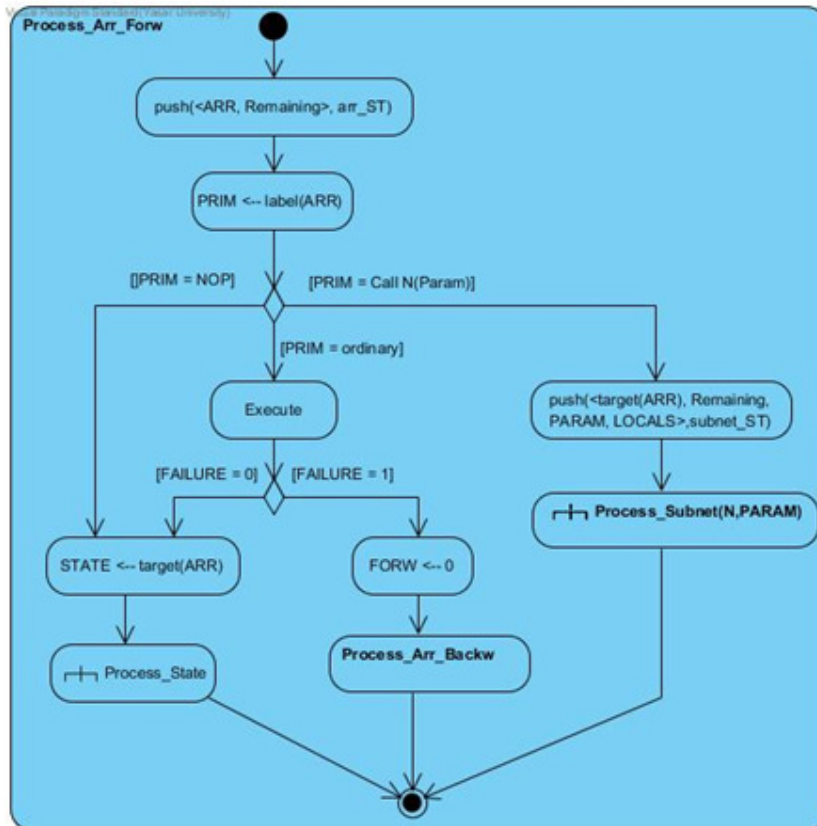


Figure 9b Activity diagram of the interpreter II

The corresponding virtual machine includes as components two stacks and a number of global variables. Stack *arr_ST* stores the current path of arrows and its elements are ordered pairs of the form $\langle ARR, Remaining \rangle$. *ARR* can take as values the name of any of the arrows in the cinnamon, as well as the special value *separator* which signals that jump to a subnet has been performed. Stack *subnet_ST* keeps information about the nested subnets invoked. Its elements have the format $\langle \text{state after the subnet call, last value of } Remaining \text{ when the call was made, values of parameters of the calling subnet, values of local variables of the calling subnet} \rangle$. The global variables are: the system flags *FORW* and *FAILURE*, the current state *STATE* of the CN, the list (stack) *Remaining* of the outgoing arrows for the current state that have not been attempted yet, current arrow *ARR*, list *PARAM* of values of the parameters of the current subnet, list *LOCALS* of the values of the local variables of the current subnet, current primitive *PRIM* to be executed. The following notations have been also used in the activity diagram: *init(SN)* is the initial node of subnet *SN*, *out(STATE)* is the list of arrows outgoing from *STATE*, and *target(ARR)* is the state which is the target of arrow *ARR*.

The interpreter presented above specifies the operational semantics of the language of cinnamons, and is a theoretical model only. Instead of an interpreter, the *SpiderCNP* IDE contains a recursive-decent-type compiler [4,11,12,22].

Note that the activity diagrams *Process_Arr_Forw* and *Process_Arr_Backw* contain actions called 'execute'. This stands for 'execute current primitive *PRIM*'. In the former activity diagram the execution is forwards, while in the latter it is performed backwards.

The backwards execution involves restoring the data. That means that normally the action primitives must have two types of action – forward action (when moving forwards) and backward

action (when moving backwards and the data are being restored). Recall that the primitives can be of two types: elementary action primitives and elementary test primitives. Only execution of action primitives (*clear*, *copy*, and *inc*) changes the environment. Execution of test primitives does not affect the environment. We can easily define a backward action for *inc*. However, there is no natural way to define a backward action to *clear* and *copy*. Technically, their backward action is empty; in reality, a programmer never uses these two action primitives in positions in the CN where a backward action might be necessary. Values of system flags may be changed by an action primitive, or directly by the control unit.

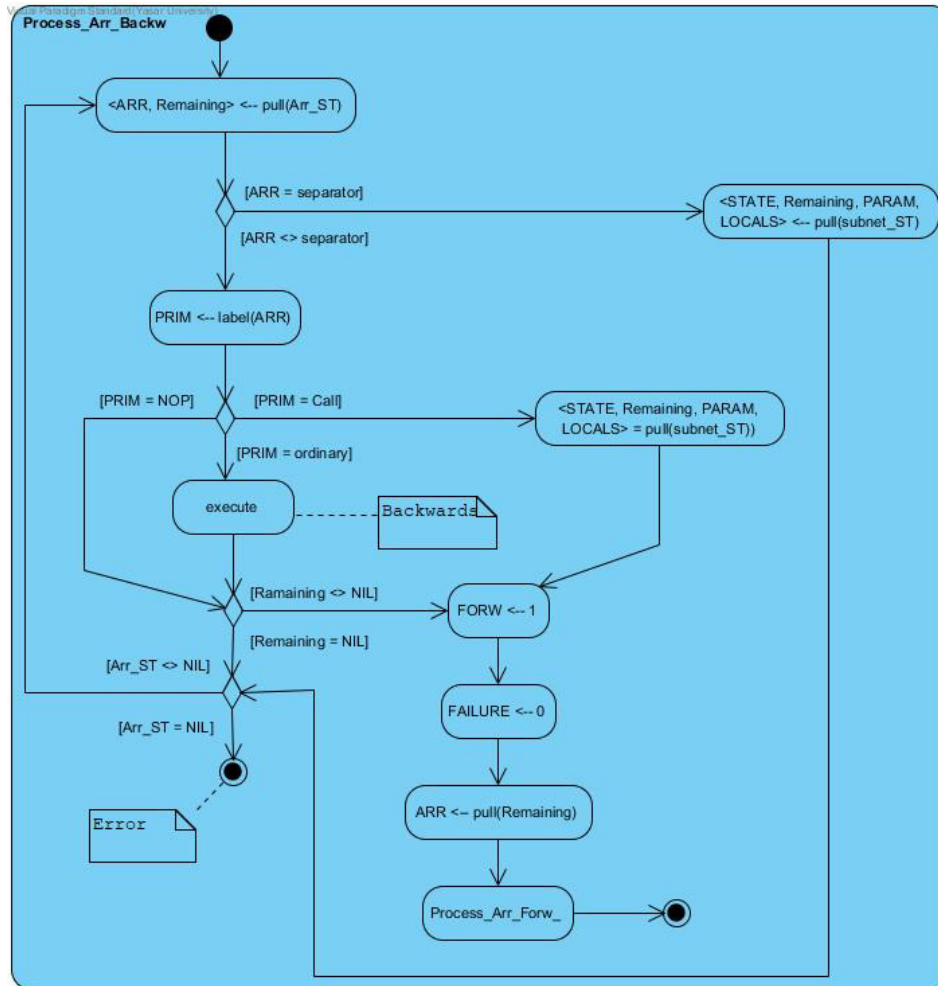


Figure 9c Activity diagram of the interpreter III

If a cinnamon is started in an initial environment ε , then in the course of the ‘computation’ the control moves in accordance with the activity diagram, and from time to time action primitives will be executed thus changing the values of variables. If and when the cinnamon halts, the final environment will in general be different from ε . We thus interpret a cinnamon Σ as a partial function $[[\Sigma]] : Env \rightarrow Env$. The value $[[\Sigma]](\varepsilon)$ is the final environment after executing the cinnamon Σ with initial environment ε , provided Σ halts. If Σ does not halt when started in initial environment ε , then $[[\Sigma]](\varepsilon)$ is undefined. Instead of saying ‘Executing the cinnamon Σ ’ one can equivalently say ‘The cinnamon Σ computes’ the function $[[\Sigma]]$. This function is the denotational semantics of Σ . Above, we use emphatic brackets $[[\]]$ following the tradition in studies on denotational semantics.

The following more formal definition of the execution of each of the allowed primitives follows elements from [13] for the case of while-programs. For $\varepsilon \in Env$, $x \in Varss$, and $a \in N$, let $\varepsilon[x \leftarrow a]$ denote the environment that is identical to ε except for the value of x , which is a . Formally,

$$\varepsilon[x \leftarrow a](y) = \varepsilon(y) \text{ if } y \neq x, \text{ and } \varepsilon[x \leftarrow a](x) = a.$$

The execution of the action primitive has the following semantics:

$$\begin{aligned} [[clear(x)]](\varepsilon) &= \begin{cases} \varepsilon[x \leftarrow 0], & \text{if } FORW = 1 \\ \varepsilon, & \text{if } FORW = 0 \end{cases} \\ [[copy(x,y)]](\varepsilon) &= \begin{cases} \varepsilon[y \leftarrow \varepsilon(x)], & \text{if } FORW = 1 \\ \varepsilon, & \text{if } FORW = 0 \end{cases} \\ [[inc(x)]](\varepsilon) &= \begin{cases} \varepsilon[x \leftarrow \varepsilon(x)+1], & \text{if } FORW = 1 \\ \varepsilon[x \leftarrow \varepsilon(x)-1], & \text{if } FORW = 0 \end{cases} \end{aligned}$$

The execution of the elementary test primitive $nonEq(x,y)$ changes the value of the system variable FAILURE only. Formally, its semantics is as follows:

$$[[noneq(x,y)]](\varepsilon) = \begin{cases} \varepsilon, & \text{if } \varepsilon(x) \neq \varepsilon(y) \text{ \& } FORW = 1 \\ \varepsilon[FAILURE \leftarrow 1], & \text{if } \varepsilon(x) = \varepsilon(y) \text{ \& } FORW = 1 \\ \varepsilon, & \text{if } FORW = 0 \end{cases}$$

As we discussed, the interpretation of a cinnamon Σ as a partial function $[[\Sigma]] : Env \rightarrow Env$. At the same time, for any given natural number j , we might want to consider a cinnamon Σ as an agent computing a j -ary partial function $f_{\Sigma} : N^j \rightarrow N$. If we want to emphasize the arity of the function, we will write $f_{\Sigma}^{(j)} : N^j \rightarrow N$.

Let a cinnamon Σ and a natural number j be given. Let n be the number of variables of the cinnamon Σ . A partial function $f_{\Sigma} : N^j \rightarrow N$ defined in the following way, is called **the partial j -ary function computed by Σ** :

- a) Suppose $j < n$. Let $\langle a_1, a_2, \dots, a_j \rangle \in N^j$. Let $\varepsilon \in Env$ be an environment such that $\varepsilon(x_i) = a_i$ for any i , $1 \leq i \leq j$, $\varepsilon(x_0) = 0$, and $\varepsilon(x_i) = 0$ for any $i > j$. Then $f_{\Sigma}(a_1, a_2, \dots, a_j)$ is defined iff $[[\Sigma]](\varepsilon)(x_0)$ is defined, and if both are defined then

$$f_{\Sigma}(a_1, a_2, \dots, a_j) = [[\Sigma]](\varepsilon)(x_0).$$

- b) Suppose $j \geq n$. Let $\varepsilon \in Env$ be an environment such that $\varepsilon(x_i) = a_i$ for any $i \leq n-1$, and $\varepsilon(x_0) = 0$. Then $f_{\Sigma}(a_1, a_2, \dots, a_j)$ is defined iff $[[\Sigma]](\varepsilon)(x_0)$ is defined, and if both are defined then

$$f_{\Sigma}(a_1, a_2, \dots, a_j) = [[\Sigma]](\varepsilon)(x_0).$$

In other words, if $j < n$ where n is the number of variables of Σ , and $\varepsilon(x_1), \varepsilon(x_2), \dots, \varepsilon(x_j)$ are the values of the j variables x_1, \dots, x_j of Σ in the starting environment while the values of the other variables are 0, then $f(\varepsilon(x_1), \varepsilon(x_2), \dots, \varepsilon(x_j))$ takes the value $[[\Sigma]](\varepsilon)(x_0)$ of the variable x_0 in the environment after the termination of the computation of Σ if it terminates. If the cinnamon has less variables than the parameters of the function then we simply use the first n variables and ignore the remaining ones.

Clearly, a given cinnamon Σ computes a nullary (constant) function $f_{\Sigma}^{(0)} : \{0\} \rightarrow N$, a unary function $f_{\Sigma}^{(1)} : N \rightarrow N$, a binary function $f_{\Sigma}^{(2)} : N^2 \rightarrow N$, and so on up to $N^{(n-1)} \rightarrow N$, and further to infinity.

As an example, let us consider the cinnamon from Fig. 3, and let $Var = \{z, x, y, \dots\}$. Let us first consider the function $f^{(2)} : N^2 \rightarrow N$. It is easy to check that $f^{(2)}(7,3) = 4$, $f^{(2)}(7,7) = 0$, $f^{(2)}(7,0) = 7$, and $f^{(2)}(3,7) = 0$. $f^{(1)}(7) = f^{(2)}(7,0) = 7$. $f^{(0)}() = f^{(2)}(0,0) = 0$. $f^{(3)}(7,3,8) = f^{(2)}(7,3) = 4$. $f^{(4)}(7,3,8,2) = f^{(2)}(7,3) = 4$.

4. TURING COMPLETENESS

In this section, we show the Turing-completeness of cinnamon programming. Of course, the power of cinnamon programming cannot be seen and appreciated when computing simple deterministic functions but when dealing with nondeterministic problems. Actually, for deterministic algorithms, cinnamon programming is a departure from the idea of structured programming. However, as cinnamon programming is another computation model, it is still worthy to show its equivalence in computational power to the other, well-established computation models which are all equivalent in power according to the famous Church-Turing thesis.

Computing models are typically considered as transducers (solvers of function problems) or acceptors (solvers of decision problems). A computation model is said to be Turing-complete if its computational power is equivalent to that of the Turing machine, and consequently to all other equivalent models. The computation model which is most similar to cinnamons, is the while-programs. Therefore, we discuss below the equivalence between cinnamons and while-programs. The result is formulated in the next theorem.

Theorem:

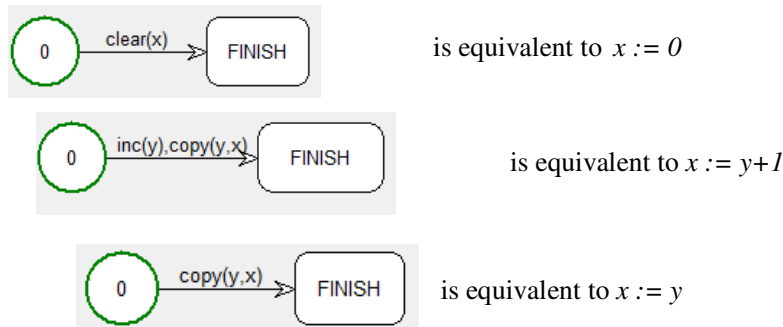
(i) For every partial function $f: N^j \rightarrow N$ computable by a while-program P , there is a cinnamon Σ such that $f_{\Sigma}^{(j)} = f$.

(ii) For every partial function $f_{\Sigma}^{(j)}: N^j \rightarrow N$ computable by a cinnamon Σ , there is a while-program P that computes $f_{\Sigma}^{(j)}$.

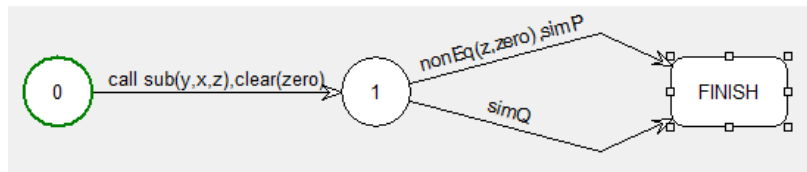
As usual for this type of results, for the proof of part (i) we need to simulate the while-program P by a suitable cinnamon, and to prove part (ii) we need to simulate the cinnamon Σ by a suitable while-program.

Slightly different (but equivalent) definitions of the concept of a while-program can be found in the literature. We will use here the definition given in [13]. This definition includes the simple assignments $x := 0$, $x := y+1$ and $x := y$, and four statement constructs: sequential composition $\{p; q\}$, conditional *if* $x < y$ *then* p *else* q , for-loop *for* y *do* p , and while-loop *while* $x < y$ *do* p . Programs built inductively from these constructs are called while-programs. The function computed by a while-program is defined similarly to that of a cinnamon.

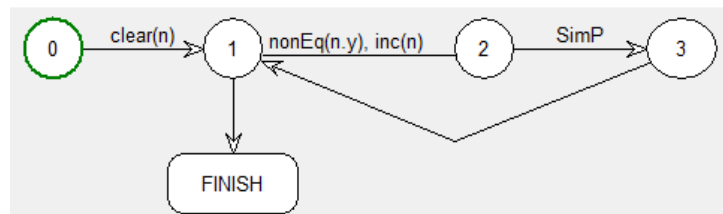
The simulation of all the constructs from the definition of a while-program by cinnamons is shown below:



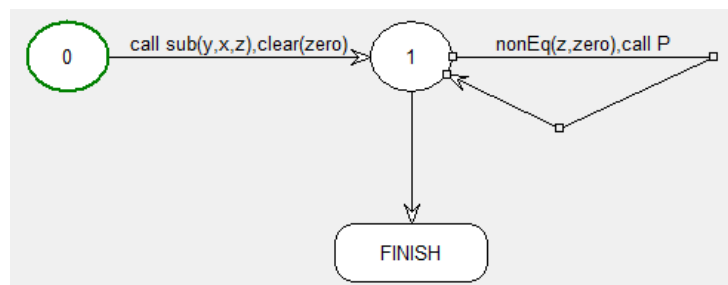
Sequential composition $\{P, Q\}$ from a while program can be simulated in the corresponding cinnamon by simulating separately P and Q , and then combining them by unifying the *FINISH* node of the part simulating P with the initial node of the part simulating Q . The simulation of the conditional statement $\text{if } x < y \text{ then } P \text{ else } Q$ is illustrated below:



where $\text{Sim}P$ is the cinnamon fragment simulating P , and $\text{sim}Q$ is the simulation of Q . A for-loop $\text{for } y \text{ do } P$ can be simulated by:



The simulation of $\text{while } x < y \text{ do } P$ is illustrated by:



where subnet P simulates segment P of the while-program.

We will not provide a formal proof of Part (ii) of the Theorem. We can notice, however, that the cinnamon interpreter presented earlier is an algorithm, and as such can be coded in a programming language or by a while-program.

5. FULL CNP VS. CINNAMON PROGRAMMING

The cinnamon, or core CNP, is a minimal theoretical computation model underlying the 'real' Control Network Programming. We include below a short list of some extensions and extra features supported by the real CNP.

- Subnets can have local variables.
- In a subnet invocation the programmer can indicate which state of the subnet will be considered as its initial state for the particular subnet call.
- The label of an arrow is a (possibly empty) sequence of primitives and/or subnet calls, not just a single primitive or a single subnet invocation.
- The user defines their own primitives which are arbitrary procedures of the underlying programming language (not only the given minimal set of three elementary action primitives and one test primitive). In addition to forward action, any user-defined primitive can also have backward action. The definitions of the primitives (plus eventually definitions of constants and helping functions) form the operational unit (in

SpiderCNP it is implemented as a file called *SpiderUnit*). Technically no difference is made between action and test primitives.

- User-defined primitives can be defined in separate units which are linked within the project.
- It should be possible that the source code of these units is written in different programming languages – interoperability.
- Object-oriented programming can be used.
- There exists a system state STOP which is equivalent to a state without outgoing arrows.
- The programmer may define costs of the arrows of the control network.
- User has very powerful control tools – control states and system options – in order to direct and control the CN traversal [24,25]. In particular, that allows simple ‘visual’ implementation of
 - heuristic algorithms [2,3,26]
 - nondeterministic algorithms [27]
 - randomized algorithms [28]
- The programmer can define the solution scope, that is, how many solutions will be found (if they exist) – a single solution, a fixed in advance number of solutions, all solutions, prompting after each solution if another solution should be sought. For example, the recognizer of arithmetic expressions (Fig. 5) will recognize as legal the expression $a+b*c$ with operation + preceding the operation *, but another solution is possible which corresponds to operation * preceding +.

6. CINNAMON PROGRAMMING, NONDETERMINISM AND OTHER EXTENSIONS

In this section, we address a few ideas for future study

As discussed in detail in [4] for the case of CNP, the concept of a cinnamon is inherently declarative and nondeterministic in nature. This comes as no surprise at all, as its skeleton is recursive nondeterministic automata – see Section 2 above. At the same time, in Section 3, we defined the cinnamon as a deterministic computation model, and then in Section 4 pointed at its Turing-completeness, that is, its equivalence in computational power to the other deterministic models of computation. It would seem natural to actually define cinnamons as a nondeterministic model (computing a relation rather than a partial function) and relate them to other nondeterministic models such as nondeterministic Turing machines, nondeterministic while-programs [29-33], etc. This direction of study is clearly related to the issue of solution scope. Further, nondeterministic computation is closely related to the definitions of a search problem, and the complexity class NP for search problems – therefore, treatment of cinnamons in this context are possible. This ‘nondeterministic’ avenue of research, however, as the other suggested topics of this section, lay beyond the scope of the current publication.

As mentioned, the ‘real’ CNP has control tools that make the ‘declarative’ implementation of randomized algorithms easy and visually clear. It would be, therefore, reasonable to appropriately extend the concept of a cinnamon and study it as a model of randomized (probabilistic) computation, in particular in relation to existing randomized (probabilistic) computation models and complexity classes [34-36].

Historically, CNP has roots in rule-based systems and fuzzy rule-based systems. Considering fuzzy cinnamons could be another possible direction for study.

REFERENCES

- [1] K. Kratchanov, E. Golemanova and T. Golemanov (2008), "Control Network Programming Illustrated: Solving Problems With Inherent Graph-Like Structure", In: Proc. 7th IEEE/ACIS Int. Conf. on Computer and Information Science (ICIS 2008), May 2008, Portland, Oregon, USA, 453-459.
- [2] K. Kratchanov, E. Golemanova, T. Golemanov and Y. Gökçen (2012), "Implementing Search Strategies in Winspider II: Declarative, Procedural, and Hybrid Approaches", In: I. Stanev and K. Grigorova (eds.): Knowledge-Based Automated Software Engineering, Cambridge Scholars Publ., 115-135.
- [3] E. Golemanova (2013), "Declarative Implementations of Search Strategies for Solving CSPs in Control Network Programming", WSEAS Transactions on Computers, 12 (4), 174-183.
- [4] K. Kratchanov, T. Golemanov, B. Yüksel and E. Golemanova (2014), "Control network programming development environments", WSEAS Transactions on Computers, 13, 645-659.
- [5] N. Jones (1997), *Computability and Complexity from a Programming Perspective*, MIT Press.
- [6] A. Kfoury, R. Moll and M. Arbib (1982, reprints 2011, 2013), *A Programming Approach to Computability*, Springer.
- [7] M. Fitting (1987), *Computability Theory, Semantics, and Logic Programming*, Oxford Univ. Press.
- [8] C. Moore and S. Martens (2011), *The Nature of Computation*, Oxford Univ. Press.
- [9] K. Slonnegger and B. Kurtz (1995), *Formal Syntax and Semantics of Programming Languages: A Laboratory Based Approach*, Addison-Wesley.
- [10] O. Goldreich (2010), *P, NP, and NP-Completeness: The Basics of Computational Complexity*, Cambridge Univ. Press.
- [11] T. Golemanov (2012), "SpiderSNP: An Integrated Environment for Visual Control Network Programming", *Annals of Ruse University*, 51, ser. 3.2, 123-127 (in Bulgarian).
- [12] T. Golemanov (2014), *Development and Study of an Integrated Development Environment for Control Network Programming*, Ph.D. Dissertation, Ruse Univ.
- [13] D. Kozen (2006), *Theory of Computation*, Springer.
- [14] W. Woods (1970), "Transition Network Grammars of Natural Language Analysis", *Comm. Of the ACM*, 13, 591-606.
- [15] E. Popov and G. Firdman (1976), *Algorithmic Foundations of Intelligent Robots and Artificial Intelligence*, Nauka (in Russian).
- [16] A. Barr and E. Feigenbaum (eds.) (1981), *The Handbook of Artificial Intelligence*, v. 1, Pitman.
- [17] K. Kratchanov (1985), *On the Foundations of Rule-Based Systems*, Dpt. Comp. Sci. Techn. Report CSTR 34/85, Brunel Univ., Uxbridge, UK.
- [18] K. Gough (1988), *Syntax Analysis and Software Tools*, Addison-Wesley.
- [19] R. Alur, M. Benedikt, K. Etessami, P. Godefroid, T. Reps, and M. Yannakakis (2005), "Analysis of recursive state machines", *ACM Trans. on Programming Languages and Systems*, 27(4):786-818.
- [20] I. Tellier (2006), "Learning Recursive Automata from Positive Examples", *RSTI – RIA – 20(2006) New Methods in Machine Learning*, 775-804.
- [21] S. Chaudhuri (2008), "Subcubic Algorithms for Recursive State Machines", <https://www.cs.rice.edu/~sc40/pubs/pop108.pdf>.
- [22] K. Kratchanov, E. Golemanova and T. Golemanov (2009), "Control Network Programs and Their Execution", In: Proc. 8th WSEAS Int. Conf. on AI, Knowledge Engineering & Data Bases (AIKED '09), Feb 2009, Cambridge, UK, 417-422.
- [23] S. LaValle (2009), "Recursive Automata", <https://courses.engr.illinois.edu/cs373/fa2009/recaut.pdf>.
- [24] K. Kratchanov, T. Golemanov and E. Golemanova (2009), "Control Network Programming: Static Search Control with System Options", In: Proc. 8th WSEAS Int. Conf. on AI, Knowledge Engineering & Data Bases (AIKED '09), Feb 2009, Cambridge, UK, 423-428.
- [25] K. Kratchanov, T. Golemanov, E. Golemanova and T. Ercan (2010), "Control Network Programming with SPIDER: Dynamic Search Control", In: *Knowledge-Based and Intelligent Information and Engineering Systems*, Proc. 14th Intl. Conf. (KES 2010), Cardiff, UK, Sep 2010, Part II, Lect. Notes in Computer Science (Lect. Notes in Artificial Intelligence), v.6277, Springer, 253-262.
- [26] K. Kratchanov, E. Golemanova, T. Golemanov and T. Ercan (2010), "Non-Procedural Implementation of Local Heuristic Search in Control Network Programming", In: *Knowledge-Based and Intelligent Information and Engineering Systems*, Proc. 14th Intl. Conf. (KES 2010), Cardiff, UK, Sep 2010, Part II, Lect. Notes in Computer Science (Lect. Notes in Artificial Intelligence), v.6277, Springer, 263-272.

- [27] K. Kratchanov, E. Golemanova, T. Golemanov and B. K ulah iođlu (2012), “Using Control Network Programming in Teaching Nondeterminism”, In: Proc. 13th Int. Conf. on Computer Systems and Technologies (CompSysTech’12), Ruse, (B. Rachev, A. Smrikarov – eds.), ACM Press, New York, 391-398.
- [28] K. Kratchanov, E. Golemanova, T. Golemanov and B. K ulah iođlu (2012), “Using Control Network Programming in Teaching Randomization”, In: Proc. Int. Conf. Electronics, Information and Communication Engineering, Macau (EICE 2012), ASME, 67-71.
- [29] E. Dijkstra (1975), “Guarded Commands, Nondeterminacy and Formal Derivations of Programs”, Comm. of the ACM, 18, 453-457. Also: E. Dijkstra. “Guarded Commands. Non-Determinacy and a Calculus for the Derivation of Programs” (EWD418), 1974
<https://www.cs.utexas.edu/users/EWD/transcriptions/EWD04xx/EWD418.html>
- [30] G. Mascari and M. Zilli (1985), “While Programs with Nondeterministic Assignments and the Logic ALNA”, Theoretical Computer Science, 40, 211-235.
- [31] J. van Leeuwen (Ed.) (1990, 1992), Handbook of Theoretical Computer Science, v. B: Formal Models and Semantics. Elsevier and MIT Press.
- [32] K. Apt, F. de Boer, E. Olderog (2010), Verification of Sequential and Concurrent Programs, 3rd ed., Springer.
- [33] K. Mamouras (2015), “Synthesis of Strategies and the Hoare Logic of Angelic Nondeterminism”, In: Foundations of Software Science and Computation Structures. 18th Int. Conf. FOSSACS 2015 Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2015, London, 11-18 Apr 2015, Proc. LNCS 9034, Springer 2015 (A. Pitts – ed.), 25-40.
- [34] R. Motwani, P. Raghavan. Randomized Algorithms. Cambridge Univ. Press, 1995.
- [35] S. Arora, B. Barak. Computational Complexity: A Modern Approach, Cambridge Uni. Press, 2009.
- [36] D. Antonova, D. Kunkle. Theory of Randomized Computation. 2005.
<http://www.ccs.neu.edu/home/kunkle/papers/techreports/randAlgo.pdf>.
- [37] K. Kratchanov (1985), Towards the Fundamentals of Fuzzy Rule-Based Systems, CSTR 35/85, Dept. Comp. Sci., Brunel Univ., Uxbridge, UK.

AUTHOR

Dr Kostadin Kratchanov taught at Ruse Univ. (Bulgaria), Technical Univ. of Sofia Plovdiv Branch (Bulgaria), Brunel Univ. (UK), European Univ. of Lefke (Northern Cyprus), Univ. of Bahrain, Grande Prairie Regional Coll. (Canada), Mount Royal Univ. (Canada), and is currently with Yasar Univ. (Turkey). His research interests are in Theoretical Computer Science, Theory of Computation, Discrete Structures, Fuzzy Automata, Applications of Category Theory in Computer Science, Programming Languages and Paradigms, Artificial Intelligence, Analysis and Design of Algorithms, Software Engineering.



REPRESENTATION WITH WORD CLOUDS AT THE PHD UNS DIGITAL LIBRARY

Georgia Kapitsaki¹ and Dragan Ivanović²

¹University of Cyprus, Cyprus

²University of Novi Sad, Serbia

ABSTRACT

Many systems provide search and recommendation capabilities to scholars that search for scientific documents including research papers and dissertations. The appearance of search results may largely affect the system use. Traditional approaches provide textual formats for showing the results to users, whereas more recent approaches concentrate on other forms, e.g., on two dimensions. Moreover, this presentation may be adapted to user needs providing a personalised user experience combined with other contextual factors, such as enriching user search with keywords from recently used documents. In this paper, we present our work on results representation in the framework of a dissertation search engine in the Serbian language with the ultimate aim to provide a more personalised experience to users. We have integrated our approach in the PhD UNS digital library system of the University of Novi Sad, a research information, library and educational information system, and are discussing an early evaluation how users are perceiving this approach outlining also our vision for a context-aware digital library system. The initial results demonstrate the usefulness of providing more choices to the users adapting application to their needs.

KEYWORDS

PHD UNS, word cloud, search results, log analysis

1. INTRODUCTION

Discovering information on the web is not always a trivial task for researchers that aim to examine previous research works that can be used as basis or reference for future research. Many systems provide the opportunity to scholars to search for papers and dissertations providing also relevant recommendations to users based on their areas of interest. Although the Google search engine is considered a superior solution to more elaborated library systems, such elaborated systems may provide more benefits in specific contexts, e.g., when searching for dissertations in specific languages, countries or institutes as addressed also in the framework of the current work [1].

Different information retrieval approaches are used in these systems for search purposes, whereas many systems integrate also recommendation mechanisms recommending relevant scientific papers or dissertations to users[2]. Although the Google Scholar recommendation is a very popular solution for scholars, additional systems may utilize different information about users in order to retrieve information. For instance, the Scienstein research paper recommender system enhances the keyword-based search by combining it with citation analysis, author analysis, source analysis, implicit ratings and explicit ratings [3].

Regardless of whether a more general or more specific system is used, the presentation of search results to users is important, as it widely affects how they perceive the system and may reduce or increase the chances of using the system and the frequency of use. In recommender systems, it is argued that the system's user interface in general (e.g., the display of predictions at the time users rate items) may even affect user's opinion [4]. The visualization may then be adapted to user needs, presenting results either in a textual or an alternative format depending on how users respond to the alternative presentations of the system. This can form part of a personalised context-aware system that considers user's environment, history and interaction with the system in order to act proactively and adapt the input and result to each user. Context-awareness is an inherent part of many systems in different domains, e.g., web services, mobile computing, where the application or system functionality adapts to the context of use [5, 6]. Context refers in most cases in any information that is relevant to the user, the system and any interaction between the user and the system[7].

Taking into consideration the above, in this paper we are presenting our work toward results representation in the framework of context-aware information provision for dissertations for scholars. In the framework of the PhD UNS digital library (DL) we are aiming at providing personalised services to the users[8]. The vision of this process is briefly described in a previous work of the authors [9]. The PhD UNS digital library includes doctoral dissertations in the Serbian language with the motivation of providing access to research data as a step toward the development of a knowledge-based society. In this paper, we focus on the presentation of the results visualization component for the search results for the users of the PhD UNS system. A new way of presenting the search results to the users was conceived, designed and implemented. Specifically, the presentation of the content of a PhD dissertation as a word cloud was addressed. Word clouds are currently widely used in different systems. A word or tag cloud is a visual representation of word content commonly used to represent content in different environments [10].

Subsequently, users have the opportunity to provide their feedback on this visualization indicating in essence whether they prefer the textual or the new graphical presentation of the results (changing from one representation to the other). The feedback was then used to adapt the results based on user preference. We are using this component as the initial step toward a fully personalised system, where different context parameters will be considered for providing a personalised context-aware user experience. At the current state, the user feedback is provided for personalisation purposes for the results appearance and is used in subsequent uses of the system. Personalisation has also been integrated into the recommendations provided by the system as presented in a previous work of the authors, whereas additional considerations are outlined as future work [11].

The contributions of this work is twofold:

- We perform a study of word clouds as a visualization approach for digital libraries search results and we evaluate this approach in the framework of the PhD UNS digital library. Although the results have been used in a digital library in the Serbian language, they can be easily replicated in libraries implemented in different languages.
- A side contribution can be found in the vision of results personalisation introduced toward a context-aware user experience.

The rest of the paper is structured as follows. Section 2 outlines previous related work in the area of recommendations for scholars and the results visualization considering also word clouds. Section 3 presents the PhD UNS digital library and its use. Section 4 is dedicated to the presentation of the new word cloud generation component including also implementation details.

The integration and use of the component in the framework of PhD UNS digital library is presented in section 5. Section 6 is dedicated to evaluation results describing how users have perceived the system and finally, section 7 concludes the paper outlining also future research directions.

2. RELATED WORK

2.1. Results visualization

Previous works have focused on personalising search or recommendation results. The process of presenting to users results in formats other than textual has been studied by many researchers in the past, in order to improve user experience in search engines, information retrieval approaches and recommender systems.

A controlled comparison of text, 2D, and 3D approaches to a set of typical information seeking tasks on a collection of 100 top ranked documents retrieved from a much larger document set was presented in [12]. The experiments conducted included the participation of 15 individuals. The study revealed that although a visualization can assist the reduction of the mental workload for interpreting the results, these reductions and their acceptance depend on an appropriate mapping among the interface, the task and the user. In relevance to the above, our approach lies in the area of 2D display of information, but instead of focusing on basic text information we have adopted newer approaches found in word clouds. Visualization has also been addressed in even earlier works in the framework of database search [13].

Most and more recent work have examined visualization in web search, such as in [14] that presents an approach for the clustering of search engine results that relies on the semantics of the retrieved documents. The approach takes into consideration both lexical and semantics similarities among documents and applies activation spreading technique, in order to generate clusters based on semantic properties. In [15], a model for web search visualization is proposed, where physical location, spatial distance, color and movement of graphical objects are used to represent the degree of relevance between a query and relevant web pages considering this way the context of users' subjects of interest. Previous works are thus trying to use different document properties in order to improve results visualization. However, we rely mostly on the document content, as it can better summarize the dissertation but focus on providing a different presentation.

2.2. Word clouds

According to Wikipedia, a word cloud is a “*visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color.*” As aforementioned, word clouds are used in different environments, whereas they are a popular way of representing information on the web summarizing the content of documents and other sources of information. Previous works have introduced various algorithms for the tag selection or new ways for the word cloud creation [16, 17, 18].

Tag clouds have been used in PubCloud for the summarization of results from queries over the PubMed database of biomedical literature[19].PubCloud responds to queries of this database with tag clouds generated from words extracted from the abstracts returned by the query. The authors found that the descriptive information is this way provided in a better way to users. However, the discovery of relations between concepts is rendered less effective. This approach has similarities

with our work, since it addresses the visualization in the framework of scientific literature, but focuses on database queries and considers only document abstracts for the tag cloud generation.

2.3. Context-awareness

Context-aware services are relevant in diverse domains. Mobile computing and pervasive computing offer the necessary information from sensors on the mobile device and in user's environments for context-aware application provision [20]. Personalisation may adapt various features (e.g., presentation, structure) in order to address specific needs of each individual [21]. In the framework of the web, and in web search and information retrieval systems, many systems utilize user's search history in order to offer personalised search. In the work of [22] it was found that personalisation based on short-term history or "within-session" behavior is less valuable than long-term or "across-session" personalisation.

In the area of digital libraries, an approach to construct personalised digital libraries that satisfy a user's necessity for information is introduced in [23]. Adaptive digital libraries are libraries that automatically learn user preferences and goals and personalise their interaction using this information. Based on this work the authors go further to develop a personalised digital library to suit the needs of different cognitive styles [24]. Adaptability versus adaptivity were investigated in a digital library and it was found that users performed better and perceived the adaptive version more positively. Both studies indicated that cognitive styles influence users' preferences for the use of digital libraries and are for this reason considered also in the framework of our work.

A number of tools have focused on providing personalised recommendations [25, 26]. However, we are not considering such works further, as the presentation of the recommendation process in PhD UNS digital library is outside the scope of the current paper.

2.4. Contribution of our work

We share similarities with previous works in terms of techniques used, as for instance word clouds have been used in other systems as well in order to improve the user experience. However, in contrast to previous works we apply a new visualization technique in a specific context, a Serbian digital library, allowing automatic adaptation for the appearance of search results based on user's reaction. This visualization concept tested in a real setting forms part of a wider context-aware experience for users.

3. PHD UNS

The current Research Information System of the University of Novi Sad (CRIS UNS) has been developed since 2008 [27]. Digital Library of Dissertations of the University of Novi Sad (PhD UNS) has been developed and integrated with the CRIS UNS system since 2011 [28, 29]. The implemented digital library enables support for the processes relevant to the educational aspect of PhD studies (release thesis for public review, promotion of PhDs, etc.). Someone can see this as an integrated system of research information system, library and educational information system. This integrated system contributes to: 1) avoiding duplicated inputs on the three platforms and thus decrease number of the university staff necessary for this work; 2) increasing metadata quality, reliability and reusability; 3) increasing quality level of services based on these metadata. PhD UNS is being used since December 2013. The University of Novi Sad Senat passed the decision that the upload of PhD thesis in PhD UNS system is the obligatory step before defending the thesis. However, all dissertations defended before December 2013 can also be scanned using Quidenus Mastered Book Scan 3.0 and be stored in the PhD UNS digital library.

Besides e-thesis, signed licenses for copyrights transfer are also stored in PhD UNS. There were two options by December of 2014: publishing dissertation under one of the six levels of Creative Commons License and publishing dissertation under a non-open-access license. Dissertations published under Creative Commons License can improve dissemination of knowledge stored in dissertations. This fact was recognized by the academic community at the University of Novi Sad. More than 90% (232 of 255) dissertations defended in the first year after putting the digital library into operation (December of 2013 – December of 2014) have been published under open-access licenses. Since December of 2014, PhD. candidates must sign open-access statement which is in accordance with the new Serbian Regulations on Higher Education.

Cataloguing within integrated system of CRIS UNS and PhD UNS is done using the MARC 21 format [30, 31]. This enables the easy integration with library information systems based on MARC formats, such as BISIS system used by Central Library of University of Novi Sad. There is a web application for searching digital library which enables search by metadata and full text of PhD dissertations [32]. Metadata for all 5,000 dissertations ever defended at the University of Novi Sad (since 1955) are stored in Digital Library of Dissertations and can be searched via web application. The search engine of the digital library has been implemented using the Apache Lucene information retrieval library [33]. The set of catalogued metadata and full text of dissertations have been pre-processed and indexed using the custom Lucene Analyzer. The Analyzer includes the following processing steps: transformation from the Cyrillic alphabet to the Latin alphabet; stop word removal based on a stop word list for the Serbian language; and stemming. More than 250,000 downloads of dissertations using the web application for searching the PhD UNS digital library have been recorded in the system logs till this point.

Also, the model of integrated system enables export of dissertation's metadata through OAI-PMH protocols in Dublin Core, MARC 21, ETD-MS and CERIF format. It enables interoperability of the PhD UNS digital library and OAI-PMH compatible institutional repositories, CERIF based information systems, MARC based library information systems, network of digital libraries based on OAI-PMH protocol, such as NDLTD, DART Europe and OATD. More than 800 open-access dissertations are exported via the OAI-PMH protocol to those networks.

4. WORD CLOUD GENERATOR COMPONENT

The word cloud generator component forms now part of the PhD UNS digital library. As aforementioned, its aim is to present user search results in a word cloud representation. Specifically, users can search for dissertations using a number of keywords as performed in similar scholar systems. The new component targets the way of presenting results to users with the aim of adapting the results visualization to the user's preferred mode, i.e., textual versus graphical. In order to achieve this, the word cloud generator component performs a number of actions on the dissertation texts as displayed in Figure 1.

The word cloud component was implemented in Java and creates as output an image (currently in PNG format) with a word cloud for the text of a PhD dissertation. The tool uses as input the PDF file of the dissertation, it then parses the textual content of the file, performs a transformation from the Cyrillic alphabet to the Latin alphabet, in order to allow the easier subsequent processing, and performs actions traditionally used in information retrieval, in order to process the document text and choose the most important words from the text : stop word removal based on a stop word list for the Serbian language and stemming. The result of pre-processing is list of pairs containing original version of word from the text and its stem. The details of the tool utilized for this pre-processing step can also be found in a previous publication [11].

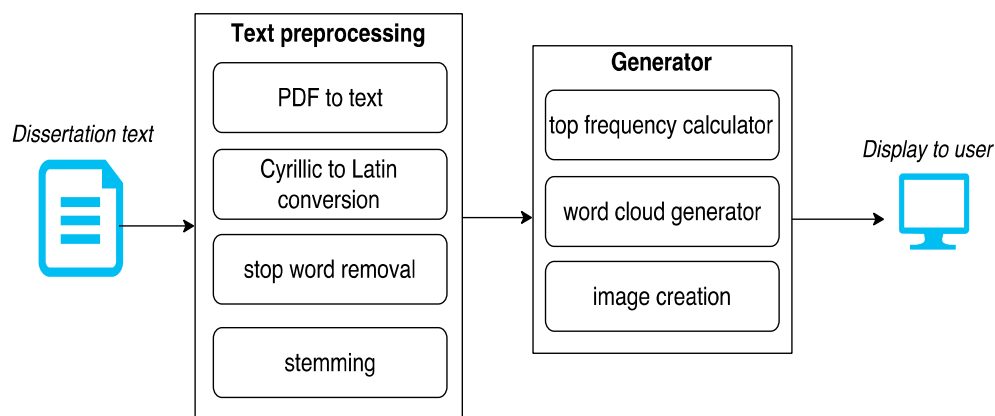


Figure 1. Word Cloud generator steps

The tool then proceeds to calculate the top frequencies of words in the text, generates the word cloud and creates an image file. All words are considered equal indetermining the importance of a keyword., while they are tokenized based on the existence of white spaces between the words in each sentence examined. The whole document text is considered instead of the abstract for instance, as we did not want to omit important words that appear in the text and may repeat many times. It is also usual for some abstracts to appear in English, but we wanted to capture the whole text. When calculating the word frequencies, the stemmed version of the words is used, in order to consider the different appearances of the same word (e.g., as noun, verb, etc.). Note however, that the words in the cloud are displayed in their original – and not the stemmed – version, in order to be better understandable to the users. The original word with the highest frequency for the respective stemmed word is presented to the user. Numbers and short words (this number of letters for small words is currently set to 4) are ignored.

For implementation purposes, the Kumo library in Java was used [34]. Kumo carries the MIT license and its code has been extended to accommodate the needs of the PhD UNS digital library. The tool can be adapted to consider a different number of keywords or use different colors for the word cloud creation.

5. INTEGRATION TO PHD UNS

The word cloud generator component described in the previous section has been integrated in the PhD UNS digital library application and has been put into operation in April, 2017. The component accepts a PDF file as input and generates an image (PNG file) as output. The information retrieval process includes two phases: indexing and searching. The purpose of creation and storing an index (indexing) is to optimize speed and performance in finding relevant documents for a search query. Without an index, searching would require considerable time and computing power. Taking into account that the word cloud generator is time-prone and a high-computing process, it is invoked in the phase of indexing and generated image is stored as supplement material to a PhD dissertation in the server file system. Figure 2 presents a Unified Modeling Language (UML) activity diagram which describes the process of adding new dissertation to the PhD UNS digital library. The activity *Generate word cloud image* is highlighted with red background and represents invoking the execution of the word cloud component described in the previous section. Moreover, the activity *Create Lucene index* includes the same steps for text pre-processing as the steps described in the word cloud generator component.

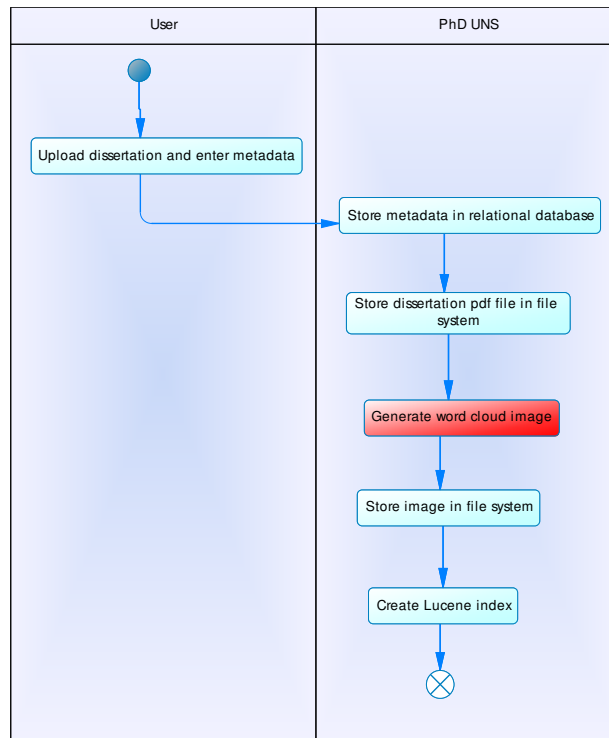


Figure 2. Adding a dissertation to the system

The default way of representing the search results to the users which access the PhD UNS search web page for the first time is randomly selected among these two choices: 1) display the results as references in Harvard representation style or 2) display the results as a word cloud image. Portions of screenshots depicting how the results are displayed in textual and word cloud representation are found in Figure 3. However, the user can request the change of the representation style as depicted in the process of Figure 4 (style option in the screenshots of Figure 3), providing this way her feedback and indicating her preference for the results visualization

6. EARLY EVALUATION

If the representation style is changed by a user of the digital library, the relevant message is stored in the server log files along with other information for the use of the system. Listing 1 shows an example of an entry in the log file with the various data stored for each user interaction with the PhD UNS system. Information, such as date and time of access, location of access, IP address, user device and representation style, are available. The representation style is also stored in cookies on the browser on the client side as the user preferred representation style and will be considered as the default visualization style for the certain user for any future access to the PhD UNS search web page. We have used the log analysis, in order to perform an evaluation of the early results of our approach and examine how users reacted to the new representation. Note that we are currently utilizing only a subset of the information available in the log files. Additional data will be used in the future in order to add more context-aware features to the digital library.

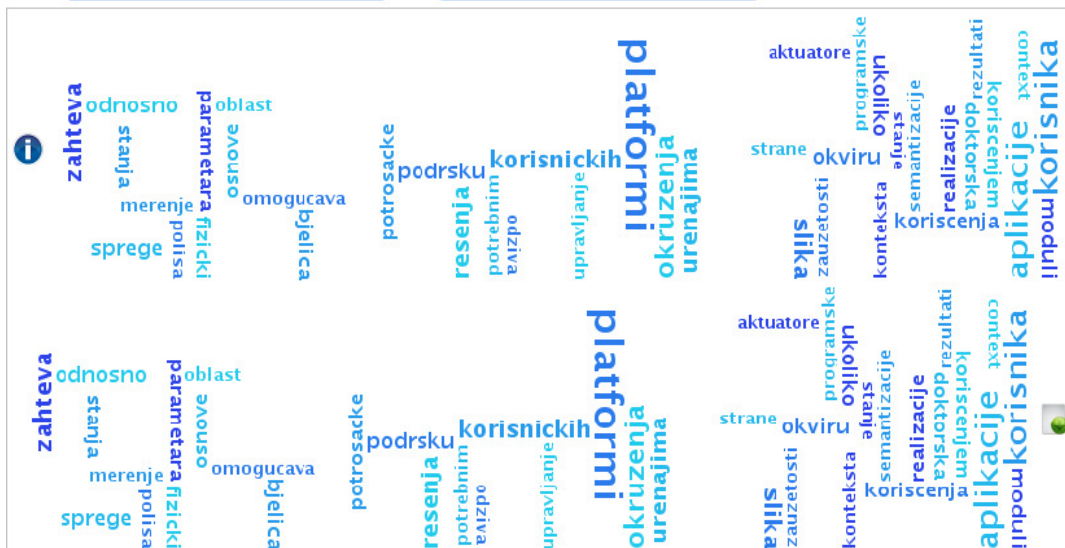
Number of results is 493

Sort by: Style:

i БЈЕЛИЦА, М. (2013) *Methods of implementation of context-aware platforms and context-aware user interfaces for applications in consumer electronics*. (PhD dissertation), Faculty of Technical Sci
 i MILANKOV, V. (2016) *Phonological awareness deficit in children with dyslexia and dysorthography*. (PhD dissertation), Faculty of Medicine at Novi Sad
 i ВΙΚΙΚИ, N. (2010) *ЋТЦАЈ ЗНАЧЕЊА ПАРТИКУЛЕ ПРОЗИРНИХ ФРАЗАЛНИХ ГЛАГОЛА НА НИЈОВО УЅВАЈАЊЕ*. (PhD dissertation), Faculty of Philosophy at Novi Sad
 i SLADIĆ, G. (2011) *Context Sensitive Access Control Model TI for Business Processes*. (PhD dissertation), Faculty of Technical Sciences at Novi Sad
 i МАТАНОВИЋ, Ј. (2016) *The significance of demographic and psychological characteristics for realistic and intended consumer behavior*. (PhD dissertation), Faculty of Philosophy at Novi Sad
 i ЕRDEJI, I. (2017) *Individual and organisational predictors of prosocial service behaviour among employees in cruise line industry*. (PhD dissertation), Faculty of Sciences at Novi Sad
 i ЛАЈШИЋ, X. (2016) *IT-Supported Development of Human Resources Management Model*. (PhD dissertation), Doktorske disertacije iz interdiscipline odnosno multidisciplinane oblasti na Unive
 i АНИЋ, I. (2011) *Cognitive processes in solving mathematical problems in real context*. (PhD dissertation), Faculty of Sciences at Novi Sad
 i ГАК, Д. (2016) *Designing Business English Course for the Purpose of Developing Pragmatic Competence Important for the Advancement of Business Communication*. (PhD dissertation), Faculty of
 i RADOVANOVIĆ, M. (2011) *High-Dimensional Data Representations and Metrics for Machine Learning and Data Mining*. (PhD dissertation), Faculty of Sciences at Novi Sad
 i NEDELJKOVIĆ, U. (2016) *Universal Type: Modernist Utopia or Current Communication Requirement*. (PhD dissertation), Faculty of Technical Sciences at Novi Sad
 i MIŠKOVIĆ, D. (2017) *Context-Dependent Speech Recognition in Human-Machine Interaction*. (PhD dissertation), Faculty of Technical Sciences at Novi Sad

(a)

Number of results is 493

Sort by: Style: 

(b)

Figure 3. PhD UNS results for search keyword “context-awareness” in (a) textual and (b) word cloud representation

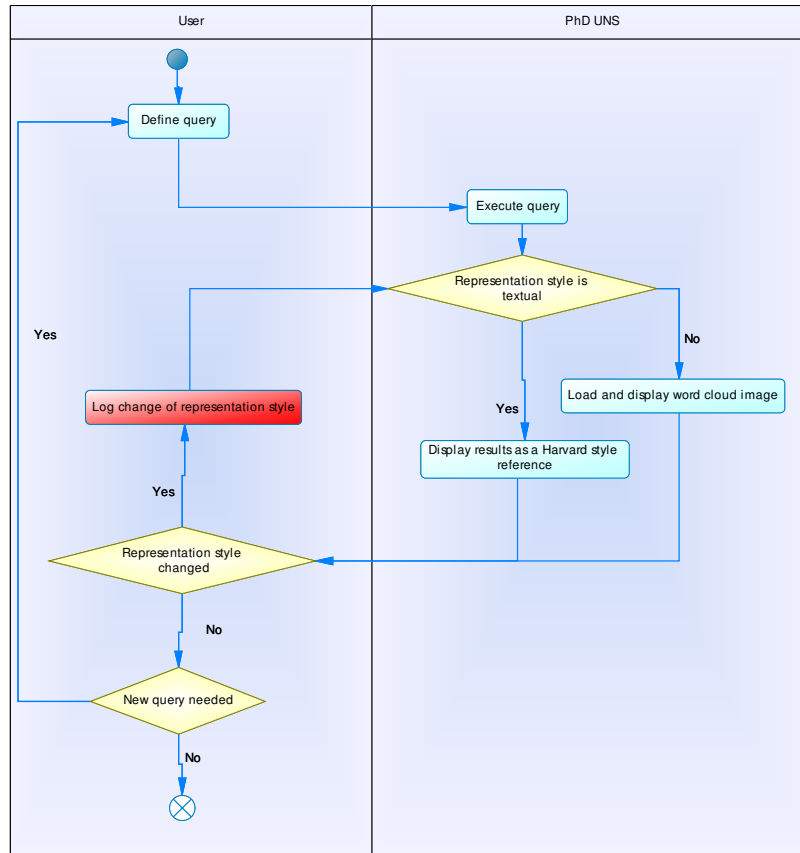


Figure 4. Query execution for representation feedback

```

[ INFO] 26.04.2017. 12:52:16 (SearchDissertationsManagedBean:setRepresentationStyle)
Date and time: Wed Apr 26 12:52:16 CEST 2017|
milliseconds: 1493203936644| +
session id: FFBAA80BECE0B982EBA313A2DE38CCC9|
userId: 149320389739914|
ip address: 147.91.177.241(proxy = 147.91.173.31)|
location: city: Belgrade, postal code: null, regionName: null (region: 00), countryName: Serbia (country
code: RS), latitude: 44.818604, longitude: 20.468094|
user agent (device): Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/49.0.2623.112 Safari/537.36|
new representation style: wordCloud
  
```

Listing 1. A log message example

Information from the users of the PhD system for the first months of use of the new representation component were collected. Log messages for the first four months of the new feature usage (April – August 2017) were for this purpose imported into a MySQL database for the purpose of log analysis. A total of 7,052 queries were defined and executed by 3,023 PhD UNS users during this period. The randomly selected representation style was changed 528 times by the PhD UNS users: it was changed 450 times to the textual representation style and 78 times to the visual (word cloud image) representation style. We can conclude that generally a lot of users accept both types of representation style taking into account that 2,495 of 3,023 users did not choose to change the representation style selected for them by the system. Furthermore, there are much more users which changed the representation style to textual than to visual representation style. This may be attributed to the fact that users are more familiar with textual

representation in their interactions with scholar systems (e.g., reference style of research papers). The provision of additional adaptability in the visualization (e.g., provide to the user the possibility to change the colors or the size of the word cloud) may also help in understanding whether the representation parameters affect the way users react to it.

However, we also observed that there are users which changed representation styles a few times; we suppose that this change was performed since they needed a different representation style for different types of information needs. We have not noted however, any decrease in the number of visitors during the period the word cloud representation as introduced, indicating that the user's frequency of use of the system and user's opinion was not largely affected. Further log analysis will be performed in the next period of use of the system, as the personalised features will expand allowing to draw more useful conclusions, where a larger number of users will be considered.

7. CONCLUSIONS

In this paper, we have presented our work on the visualization of recommendations for scholars in the framework of the PhD UNS digital library. We have used a word cloud in order to study how users react to this new representation and then adapt the default results presentation for each user based on the feedback received by users. To the best of our knowledge, this is the first work that addresses the study of results representation for the Cyrillic alphabet. The initial results obtained from analyzing the log files of the system demonstrate that most users accept both representation styles, as only a small percentage of users chose to change the randomly chosen representation style during the use of the system. Further evaluation is required, in order to study whether the adaptability of this word cloud representation can improve the way users interact with PhD UNS.

As future work, we intend to extend the personalised features of the system providing context features to the user searchers via the enrichment of user queries and results for scientific documents retrieval with context information. We intend to use different information sources for this purpose, such as keywords from dissertations texts, user device motion information (when a mobile device is used to access the system), user device battery information, keywords from previous user searches, and keywords from user's publications and the publications of user's collaborators (available only for registered users). These adaptations will realize our vision for a personalised digital library system for the Serbian language. Note that we want to preserve also user's privacy at the time of providing a personalised solution. Although all user data is currently anonymous and most visitors are guest users of the system, we will target in the future additional privacy protection mechanisms for registered users.

ACKNOWLEDGEMENTS

The work is partially supported by the Information and Communication Technologies (ICT) COST Action IC1302: "KEYSTONE - semantic KEYword-based Search on sTructured data sOurcEs".

REFERENCES

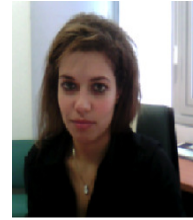
- [1] Brophy, Jan, & David Bawden (2005) "Is Google enough? Comparison of an internet search engine with academic library resources", *Aslib Proceedings*. Vol. 57. No. 6. Emerald Group Publishing Limited.
- [2] Beel, Joeran, Langer, Stefan, Genzmehr, Marcel & Nuernberger. Andreas (2013) "Introducing Docear's research paper recommender system", In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries (JCDL '13)*, ACM, pp459-460.

- [3] Gipp, Bela, Beel, Joeran & Hentschel, Christian (2009) "Scienstein: A research paper recommender system." Proceedings of the international conference on emerging trends in computing (icetic'09).
- [4] Cosley, Dan, Shyong K. Lam, Istvan Albert, Joseph A. Konstan & John Riedl (2003) "Is seeing believing?: how recommender system interfaces affect users' opinions", In Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, pp585-592.
- [5] Hong, Jong-yi, Suh, Eui-ho & Kim, Sung-Jin (2009) "Context-aware systems: A literature review and classification." Expert Systems with Applications, Vol. 36, No. 4, pp8509-8522.
- [6] Kapitsaki, Georgia M., Prezerakos, George N., Tselikas, Nikolaos D. & Venieris, Iakovos. S. (2009) "Context-aware service engineering: A survey", Journal of Systems and Software, Vol. 82, No. 8, pp1285-1297.
- [7] Abowd, Gregory, Dey, Anind, Brown, Peter, Davies, Nigel, Smith, Mark & Steggles, Pete (1999) "Towards a better understanding of context and context-awareness", In Handheld and ubiquitous computing, Springer Berlin/Heidelberg, pp304-307.
- [8] Borgman, Christine L. (1999) "What are digital libraries? Competing visions", Inf. Process. Manage., Vol. 35, No. 3, pp227-243.
- [9] Ivanovic, Dragan, & Kapitsaki, Georgia M., (2015) "Personalisation of Keyword-based Search on Structured Data Sources", 1st International KEYSTONE Conference (IKC 2015).
- [10] Scansfeld, Daniel, Vanessa Scansfeld, & Elaine L. Larson, (2010) "Dissemination of health information through social networks: Twitter and antibiotics", American journal of infection control, Vol. 38, No. 3, pp182-188.
- [11] Azzopardi, Joel, Ivanovic, Dragan, Kapitsaki, Georgia M. (2016) "Comparison of Collaborative and Content-Based Automatic Recommendation Approaches in a Digital Library of Serbian PhD Dissertations", Proceedings of the International KEYSTONE Conference 2016, pp100-111.
- [12] Marc M. Sebrecths, John V. Cugini, Sharon J. Laskowski, Joanna Vasilakis, & Michael S. Miller. (1999) "Visualization of search results: a comparative evaluation of text, 2D, and 3D interfaces", In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99), ACM, pp3-10.
- [13] Bowers, Frank H., and Stuart K. Card, (1996) "Method and apparatus for visualization of database search results", U.S. Patent No. 5, pp546,529.
- [14] Soliman, Sara Saad, El-Sayed, Maged F. & Hassan, Yasser F. (2015) "Semantic Clustering of Search Engine Results", The Scientific World Journal.
- [15] Nguyen, Tien, & Zhang, Jin (2006) "A novel visualization model for web search results", IEEE transactions on visualization and computer graphics, Vol. 12, No. 5.
- [16] McNaught, Carmel & Lam, Paul (2010) "Using Wordle as a supplementary research tool", The qualitative report, Vol. 15, No. 3, pp630.
- [17] Cui, Weiwei, Yingcai Wu, Shixia Liu, Furu Wei, Michelle X. Zhou, & Huamin Qu, (2010) "Context preserving dynamic word cloud visualization", In IEEE Pacific Visualization Symposium (PacificVis), pp121-128.
- [18] Hassan-Montero, Yusef & Herrero-Solana, Victor (2006) "Improving tag-clouds as visual information retrieval interfaces", In Proceedings of the International conference on multidisciplinary information sciences and technologies, pp25-28.

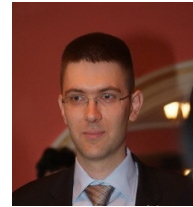
- [19] Byron Y-L Kuo, Thomas Hentrich, Benjamin M. Good & Mark D. Wilkinson. (2007) "Tag clouds for summarizing web search results", In Proceedings of the 16th international conference on World Wide Web (WWW '07). ACM, pp1203-1204.
- [20] Raento, M., Oulasvirta, A., Petit, R., & Toivonen, H. (2005) "Context Phone: A prototyping platform for context-aware mobile applications", IEEE pervasive computing, Vol. 4, No. 2, pp51-59.
- [21] Fink, Josef & Kobsa, Alfred (2000) "A review and analysis of commercial user modeling servers for personalisation on the world wide web", User Modeling and User-Adapted Interaction, Vol. 10, No. 2, pp209-249.
- [22] Yoganarasimhan, Hema, (2016) "Search personalisation using machine learning".
- [23] Frias-Martinez, E., Magoulas, G., Chen, S. & Macredie, R., (2006) "Automated user modeling for personalised digital libraries", International Journal of Information Management, Vol. 26, No. 3, pp234-248.
- [24] Frias-Martinez, Enrique, Sherry Y. Chen, & Xiaohui Liu, (2009), "Evaluation of a personalised digital library based on cognitive styles: Adaptivity vs. adaptability", International Journal of Information Management, Vol., 29, No. 1, pp48-56.
- [25] Tejada-Lorente, Álvaro, Carlos Porcel, Eduardo Peis, Rosa Sanz, & Enrique Herrera-Viedma, (2014) "A quality based recommender system to disseminate information in a university digital library", Information Sciences, Vol. 261, pp52-69.
- [26] Beel, Joeran, Langer, Stefan, Kapitsaki, Georgia, Breiting, Corinna & Gipp, Bela (2015) "Exploring the potential of user modeling based on mind maps", In Proceedings of the International Conference on User Modeling, Adaptation, and Personalisation, Springer, pp3-17.
- [27] Surla, Dušan, Ivanović, Dragan & Konjović, Zora (2013) "Development of the software system CRIS UNS." In Proceedings of the Intelligent Systems and Informatics (SISY), 2013 IEEE 11th International Symposium, pp111-116.
- [28] CRIS PhD UNS, <http://dosird.uns.ac.rs/phd-uns-digital-library-phd-dissertations>.
- [29] Ivanović, Lidija, Ivanović, Dragan & Surla, Dušan (2012) "Notes on Operations: Integration of a Research Management System and an OAI-PMH Compatible ETDs Repository at the University of Novi Sad, Republic of Serbia", Library Resources & Technical Services, Vol.56, No. 2, pp 104-112.
- [30] Ivanović, Dragan, Surla, Dušan & Konjović, Zora (2011) "CERIF compatible data model based on MARC 21 format", The Electronic Library, Vol. 29, No. 1, pp52-70, DOI: 10.1108/02640471111111433.
- [31] Ivanović, Lidija, Ivanović, Dragan & Surla, Dušan (2012) "A data model of theses and dissertations compatible with CERIF, Dublin Core and EDT-MS", Online Information Review, Vol. 36, No. 4, pp568-586, DOI 10.1108/14684521211254068.
- [32] Ivanović, Lidija, Ivanović, Dragan, Surla, Dušan & Konjović, Zora (2013) "User interface of web application for searching PhD dissertations of the University of Novi Sad", In Proceedings of the Intelligent Systems and Informatics (SISY), 2013 IEEE 11th International Symposium, pp117-122.
- [33] Apache Lucene, <https://lucene.apache.org/>.
- [34] Kumo, <https://github.com/kennycason/kumo>.

AUTHORS

Georgia Kapitsaki is an Assistant Professor at the Department of Computer Science of the University of Cyprus and faculty member of the Software Engineering and Internet Technologies (SEIT) laboratory in UCY. She received her PhD from the National Technical University of Athens, Greece (2009). Her research interests include: software engineering, privacy enhancing technologies (PETs), context-awareness and service-oriented computing. She has published over 40 papers in international conferences and journals. She has participated in conference organisation (e.g. ICSR 2016) and has served as a TPC member in repudiated journals and conferences. She has been involved in EU projects and has worked as a software engineer in the industry.



Dragan Ivanović is an associate professor within Faculty of Technical Sciences, University of Novi Sad. His research interests are focused on digital documents management, information retrieval, research management information systems, software architectures and library standards. Dragan holds courses Digital documents management on master studies and Selected Topics of Digital Archives on PhD studies at the Computer Science Department – these courses include topics related to: library standards and formats, digital repositories (including OA policies and OA repositories), information retrieval, search personalisation, etc. Dragan has been involved in more than 10 research projects and published over 40 scientific papers.



INTENTIONAL BLANK

ATTACK ANALYSIS IN VEHICULAR AD HOC NETWORKS

Ömer Mintemur and Sevil Sen

Computer Engineering, Hacettepe University, Ankara, Turkey

ABSTRACT

One of the most promising and exciting areas of communication technology is Vehicular Ad Hoc Networks (VANETs). It enables vehicles to communicate among and between each other and fixed infrastructures, and, to provide a safe and enjoyable driving experience. However, VANETs are very susceptible to attacks that could easily be evasive due to its dynamic topology, and, resulting in very dramatic results in traffic. To develop a suitable security solution for VANETs, it must first be understand how such attacks could affect the network. Therefore, this study analyzes four different types of attacks against two popular routing protocols (AODV, GPSR) in VANETs. All attacks, blackhole, dropping, flooding, and bogus information, were implemented on two real maps having low and high density. The results clearly show how attacks could severely affect communication and, the need for security solutions for such highly dynamic networks.

KEYWORDS

VANETs, security, attacks, blackhole, bogus, information.

1. INTRODUCTION

Conventional communication technology is changing rapidly. The opportunity to communicate via wireless technology brings about unlimited alternatives such as mobile ad hoc networks (MANETs), and wireless sensor networks (WSN). In mobile ad hoc networks, mobile nodes can communicate with no fixed infrastructure. This infrastructureless characteristic of mobile ad hoc networks enables the application of many different communication technologies. One of the most intriguing is vehicular ad hoc networks (VANETs). Basically, this new environment enables communication among and between vehicles and fixed structures called Road Side Units (RSUs). In such networks, each vehicle is equipped with a device called an On-Board Unit (OBUs) that enables their communication capability [1]. Vehicles can send and receive information such as traffic conditions and, road conditions [2]. The main purpose of VANETs is to provide drivers with a safer and more efficient driving experience. VANETs are expected to become widespread once certain research challenges have been successfully addressed, such as provision of security for these dynamic networks.

Although VANETs are highly desirable for a safe and comfortable driving experience, the use of wireless channels and fast changing topology make them vulnerable to new forms of attack [3]. A malicious vehicle could disrupt the network and, cause unwanted results such as loss of lives, money, and time [3], [4]. An attacker could achieve its purpose mainly through exploitation of the weakness of the routing protocols and application protocols in VANETs.

An extensive analysis of attacks is necessary in order to develop suitable security solutions for VANETs, which is the primary aim of this study. In this study, four types of attacks, namely

blackhole, dropping, flooding, and bogus information attacks were analyzed on two popular routing protocols, AODV and GPSR. Real high/low density road maps were simulated in which vehicles move as on real roads. Furthermore, attack scenarios were implemented on real maps having realistic conditions (network mobility and density). The code and configuration files of attack simulations will be made publicly available. The authors believe that this analysis helps researchers to create efficient and suitable security solutions for VANETs.

2. RELATED WORK

Analysis of attacks in AODV have been widely analyzed in the literature. However, such analyses are mostly conducted out on mobile ad hoc networks, rather than, highly dynamic vehicular ad hoc networks. Furthermore, there has been little study of attacks in GPSR on VANETs.

Extensive analysis of different types of attacks against AODV on MANETs can be found in [5]. In this current study, both atomic and compound misuses were introduced for AODV. In the simulations, only one attacker was assumed to be in the network. Furthermore, the simulated networks consisted of only five nodes in atomic misuses, and 20 nodes in compound misuses. Even though this study presents all kinds of attacks in detail, the simulations were limited.

One of the mostly analyzed attacks to be found in the literature is the blackhole attack, due to being a specific attack to ad-hoc routing protocols. Four routing protocols (AODV, DSR, OLSR and TORA) were analyzed under blackhole attack in MANETs [6]. The results showed that AODV performed poorer than other protocols on simulated networks under attack. Blackhole attack was also analyzed in VANETs by using AODV and OLSR [7]. The results support the study given in [6] that AODV is more susceptible to attacks than OLSR. Although the simulations were for VANETs, the nodes in the experiments were assumed to move at a constant speed (10 m/s), which is unrealistic for vehicular communication.

As in MANETs, the watchdog-based detection mechanism is usually proposed for the detection of blackhole attacks in VANETs [8]. With this method, every packet sent by vehicles is watched. Each vehicle maintains a trust table for its neighbors, and the trust value is determined by the ratio of packets that should be transmitted over packets actually transmitted. Any vehicle that drops below a certain threshold is considered malicious.

In the literature flooding attack [9] is another type of attack analyzed for MANETs, where network performance is greatly affected by the sending of numerous packets [10]. This current study also used AODV as an exemplar protocol. The current study also proposed a detection mechanism for ad hoc flooding attack in which every vehicle watches its neighbors. If a neighbor sends RREQ packets exceeding a certain threshold, it is tagged as an attacker. A similar threshold-based approach [11] is proposed for the detection of flooding attacks on VANETs. For further information on attack detection mechanisms in VANETs, see the recent surveys in this area [12],[13].

As shown in the literature, analysis of attacks on VANETs is very limited. Moreover, although a bogus information attack could have a disastrous effect on VANETs, the literature mainly proposes a detection technique, and does not analyze the attack in detail as in this current study. Furthermore, the simulation environment in some studies might be unrealistic. In this current study, real high/low density road maps are simulated in which vehicles move as on a real road. To the best of the authors knowledge, this current study is the most extensive attack analysis in terms of attacks type, and the number of attackers in VANETs.

3. ROUTING PROTOCOLS: AODV AND GPSR

VANETs can inherit routing protocols currently used in MANETs. An extensive review of routing protocols of VANETs can be found in [14]. This current study employs widely known AODV (Ad-Hoc on Demand Distance Vector Routing) [15] and GPSR (Greedy Perimeter Stateless Routing) [16] routing protocols. This section briefly explains these two routing protocols. While AODV is one of the most popular routing protocols, GPSR is one of the position-based protocols suited to VANETs [17].

3.1. AODV (Ad-Hoc on Demand Distance Vector Routing Protocol)

AODV routing protocol is a reactive routing protocol [15] in which the routes are established just before any packet transmission begins. In the route discovery, two types of routing control packets are used: RREQ (route request) and RREP (route reply).

When a vehicle wants to send a data packet to another vehicle and do not know the path to this destination vehicle, a RREQ packet is generated and broadcast to the network. Vehicles that receive these RREQ packets check their routing table as to whether or not they already know a path to the destination vehicle. If they locate a fresh route to the destination vehicle, they return a RREP packet to the source vehicle. Otherwise, the RREQ packet is rebroadcast. When a RREQ packet arrives to the destination, a unicast RREP packet is returned to the source vehicle. As soon as the source node receives a RREP packet, it starts sending data packets. There could be more than one path between two communication endpoints, but the shortest path is built in AODV.

AODV also has a routing control packet called RERR (Route Error), which are sent by vehicles if any of their neighbors are unreachable. This packet type indicates broken links, vehicles that have gone out of range, etc. The local connectivity could be maintained both at the link layer and at the routing layer. If a link breakage is detected, RERR packets are sent to the neighbors.

3.2. GPSR (Greedy Perimeter Stateless Routing Protocol)

GPSR routing protocol is a geographically-based routing protocol which transmits data packets by using vehicles' geographical positions [16]. Unlike AODV, GPSR does not establish a route in advance.

GPSR uses two different forwarding mechanisms: greedy and perimeter forwarding. In GPSR, vehicles know their neighbors by sending periodic beacon packets. Through the sending and receiving of beacons, vehicles each construct their own routing table. At the beginning, positions of each vehicle are saved in a look up table. When a vehicle moves, the look-up table is updated with the new position of the vehicle by using LocService (LOCS) packets which are periodic packets informing about vehicles' positions. When a vehicle wants to send a message, it originates a packet containing only the originator address and the destination address. The source vehicle transmits the packet to its neighbor closest to the destination, according to the neighbors' positions. This mechanism continues until the destination is reached (greedy forwarding). Hence, the next hop is determined by forwarding nodes during data packet transmission. When greedy forwarding fails, it means the packet transmitting vehicle cannot find any vehicle closer to the destination within its coverage area; hence GPSR turns to perimeter forwarding. In perimeter forwarding, packets are forwarded using the planar graph. Packets are traversed by the right hand rule within the network until the packet transmission turns back to greedy forwarding. As stated in [16], beacon intervals could be selected optionally. In the current study, the beacon interval was selected as 0.5 s to ensure compatibility with the nature of VANETs. The literature shows that the bigger the beacon interval, the fewer packets are delivered successfully [16].

Hierarchical location service [18], which divides the area covered by the network into a hierarchy of regions for discovering the locations of nodes, is also employed in the simulations.

4. IMPLEMENTED ATTACKS

In this current study, the effects of four types of attacks were evaluated on both routing protocols. The implementation details of these attacks on AODV and GPSR are detailed in this section.

4.1. Blackhole Attack

The main aim of this attack is to direct data packets to the malicious vehicle by claiming it has the best route to the destination. It is mainly employed with dropping attack. After the route is established through the malicious vehicle, data packets are dropped.

In AODV, the freshness of a route is defined with sequence numbers. In the blackhole attack scenario of the current study, the attacker takes advantage of this characteristic of AODV. The malicious vehicle receiving a RREQ packet replies with a RREP packet by incrementing the destination sequence number in the original RREQ packet. Even though the source node could receive more than one RREP packet, it will accept the freshest one coming from the malicious vehicle. Hence the malicious vehicle place itself in the route between the source and the destination node. The malicious vehicle could either listen to or disrupt the source vehicles' communication. In this attack scenario, the attacker simply drops data packets it receives.

In GPSR, the source vehicle always chooses a vehicle closest to the destination for forwarding its packet. In this attack scenario, the attacker takes control of the traffic by advertising itself as the nearest node to the destination. As in AODV, the malicious vehicle drops data packets it receives. In order to achieve its goals, the attacker needs to be accessible to the source node in order to receive the request and send a fake reply.

4.2. Dropping Attack

In this attack type, the malicious vehicle simply drops all the packets it receives. This attack is different from a blackhole attack. In the blackhole attack scenario, the malicious vehicle claims itself to have the shortest path and takes control of the traffic, then drops the data packets. However, in a packet dropping attack scenario, the malicious vehicle only drops data packets if a packet is transmitted through it. Even a simple dropping attack could cause serious consequences, especially in safety-related applications. Furthermore, it is difficult to distinguish from legal packet dropping on networks with high mobility.

4.3. Flooding Attack

The flooding attack is a type of DoS attack. The main aim of the attack is to exhaust the network by sending numerous control packets, resulting in network nodes unable to process legitimate traffic. While malicious vehicles could bombard the network with RREQ packets in AODV, beacon messages are employed in GPSR for this purpose. This attack both exhausts network bandwidth and nodes' packet queues, and the network becomes unavailable to legitimate users.

In the current study's simulations, in AODV a malicious vehicle broadcasts a fake RREQ packet for a non-existent vehicle in the network every 0.2 seconds. In GPSR, a malicious vehicle broadcasts lots of beacons to its neighbors in order to disrupt their functionalities. Beacon packets are sent at 0.2 second intervals. Fake packets are continually sent in both routing protocols until the simulation terminates.

4.4. Bogus Information Attack

In bogus information attacks, the attacker sends falsified information to the network. For example, an attacker could send information about a fake road accident in order to divert traffic onto another road. This scenario could be very effective when there is no other vehicle to verify this deception of the falsified information. It is termed as a motorway attacker [19] if the attacker moves around quickly, and disseminates false information to a large group of nodes.

In the attack scenario, the attacker chooses a node as its victim, and then prepares a RREQ or beacon packet for AODV and GPSR respectively as generated from the victim. The packets are generated for a randomly selected destination node, and the attacker node broadcasts these packets on behalf of the victim node every five seconds. The attacker attracts traffic by being the freshest node or the closest node to the destination in AODV and GPSR respectively. Again, any packets transmitting through the attacker will be dropped. This attack could also be used to isolate a node from the network; however, it will have little effect on the network due to the fast changing topology of VANETs. Packets not transmitted through the attacker will remain unaffected.

5. EXPERIMENTAL RESULTS

In this section, firstly the simulation environment is introduced. Then, the effects of each attack on the network are evaluated by analyzing simulation results. Each attack is evaluated against well-known network performance metrics: packet delivery ratio, overhead, end-to-end (E2E) delay.

5.1. Simulation Environment

All simulations are conducted in a widely used network simulator, ns-2 [20]. Each simulation is run for a period of 200 seconds. Each attack is evaluated in networks with varying numbers of attackers (0%, 5%, 10%, 15%, 20%, 25%, and 30%). In each group of attackers, the position of attackers is assigned randomly 10 times. 10 different connection files are established, and each connection file has 15 different connections. Hence, 700 simulations are run for an attack against a routing protocol, and their averaged results are presented in the subsequent section. In total, 5,600 simulations are ran for a map. The simulation parameters used in the experiments are given in Table 1.

Table 1: Simulation Parameters

| Simulation Parameters | Value |
|-------------------------|--|
| Simulation Time | 200 seconds |
| Network Area | Istanbul Highway (2600m X 1340m) Munich City Center (2000m X 1380m) |
| Number of Vehicles | 35 |
| Data Packet Type | CBR |
| Packet Size | 512 bytes |
| Vehicle Speed | 0 – 70 m/s |
| Propagation Model | Nakagami [21] |
| Communication Range | 250 m |
| MAC Layer Protocol | 802.11 |
| Local Link Connectivity | Link Layer Notifications (MAC Control Packets) |

Simulations are implemented on two real maps: Munich city center, and a part of the Istanbul Highway network. These roads were chosen due to their traffic densities. While the Munich road has high density, the Istanbul Highway has low density. These maps are generated by using SUMO [22] and OpenStreetMap [23].

5.2. Results in AODV

5.2.1. Packet Delivery Ratio – AODV

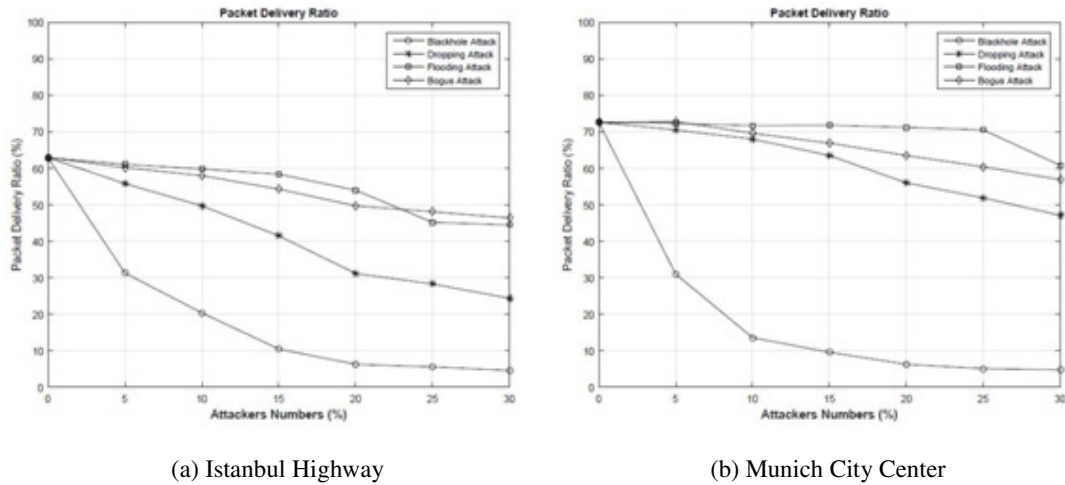


Figure 1. Packet Delivery Ratio – AODV

Figure 1 shows the packet delivery ratio of AODV in Istanbul Highway 1(a) and Munich City Center 1(b). In general, a dense network has a higher packet delivery ratio than a sparse network. As expected, while the attacker percentage in the network increases, packet delivery ratio decreases in both maps. Figure 1 clearly shows that the Istanbul Highway is affected more severely than Munich City Center. Because of the density, vehicles in Munich are able to find more connections than Istanbul Highway even with existence of attackers.

Packet dropping attack decreases the packet delivery ratio as expected; however, the increase is not as much as in the blackhole attack scenario. This attack is more effective if the attacker is in a critical position such as being the only node that connects two endpoints, or two network partitions [24]. Since the attacker diverts traffic through itself in a blackhole attack, it is more effective. However in a simple packet dropping attack scenario, the attacker only drops packets if they are transmitted through it.

Flooding attack does not have as severe effect as blackhole and dropping attacks do. As the number of fake packets broadcast to the network increases, it will cause more packets to be dropped due to heavy traffic impacting the network. This situation applies to the increase of the number of attackers as clearly seen in the figure 1.

In the bogus attack scenario, by pro-actively forging fake routing control packets without receiving any packets (differently from a blackhole attack), the attacker diverts and then drops data packets, and hence decreases the packet delivery ratio as shown in Figure 1.

In general, sparse networks (Istanbul Highway) are affected more than dense networks (Munich City Center). Moreover, as expected, when there are no malicious vehicles in the network, dense

networks have a higher packet delivery ratio than dense networks. In such networks, vehicles can find more vehicles able to continue the packet transmission.

5.2.2. Overhead – AODV

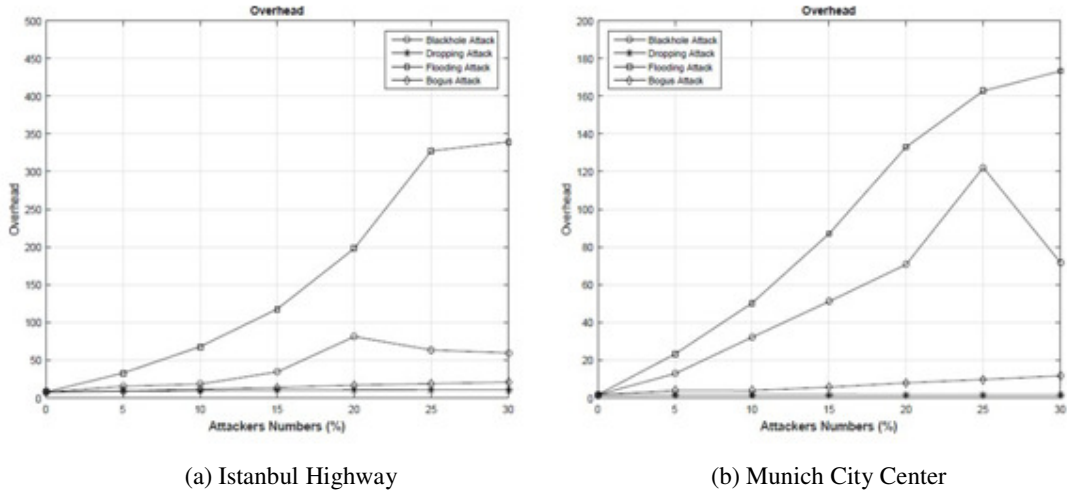


Figure 2. Overhead – AODV

Figure 2 shows the overhead results for the attacks in both Istanbul Highway and Munich City Center. As the number of attacker increases, the overhead also increases due to disrupted routes. Flooding attack due to its very nature increases overhead the most. Blackhole attack also increases the overhead considerably due to its disruption of effective routes. The density of maps affects the overhead results as well. Since the dense network provides more connectivity, less control packets are introduced to the network.

5.2.3. End-to-End Delay – AODV

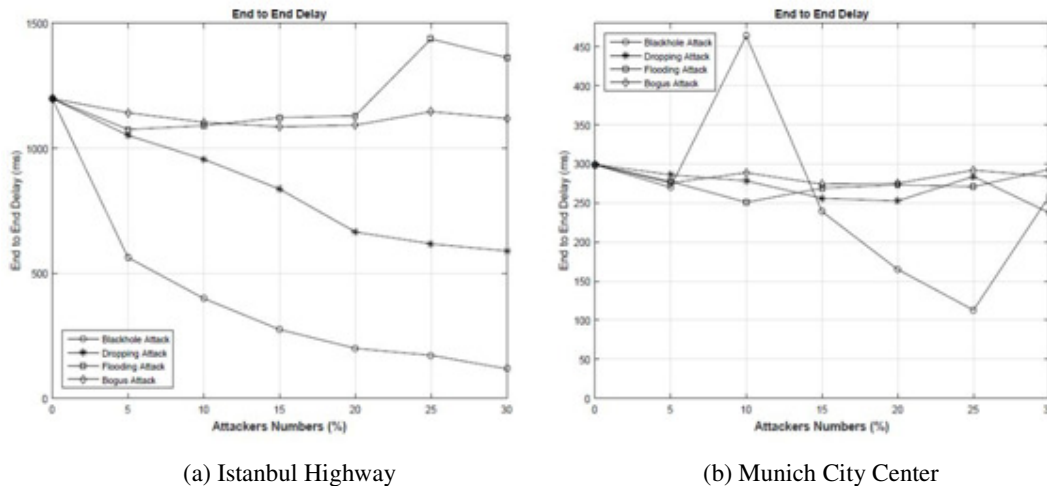


Figure 3. End - to - End Delay - AODV

Istanbul Highway is affected much more than Munich City Center in terms of end-to-end delay as shown in Figure 3. End-to-end delay remains the same or increases when the number of attackers

exceeds a certain threshold in flooding and bogus information attacks. In the existence of blackhole or dropping attacks, since less data packets are trying to be sent, they will be able to reach their destinations without waiting due to traffic levels in the network. Even though the number of routing control packets increases, as shown in Figure 2, the increase is not very significant. Because of dropped data packets, routes to the destination are re-built. In the simulations, it is observed that the average hop count could also decrease while the number of attackers increases and the topology changes. Due to sending data packets to closer nodes, a decrease in end-to-end delay also occurred in the case of blackhole and dropping attacks. There was a fluctuation seen in the blackhole attack in Munich City map in Figure 3, probably caused by the selection of attackers, position of attackers, communication patterns, etc.

5.3. Results in GPSR

5.3.1. Packet Delivery Ratio – GPSR

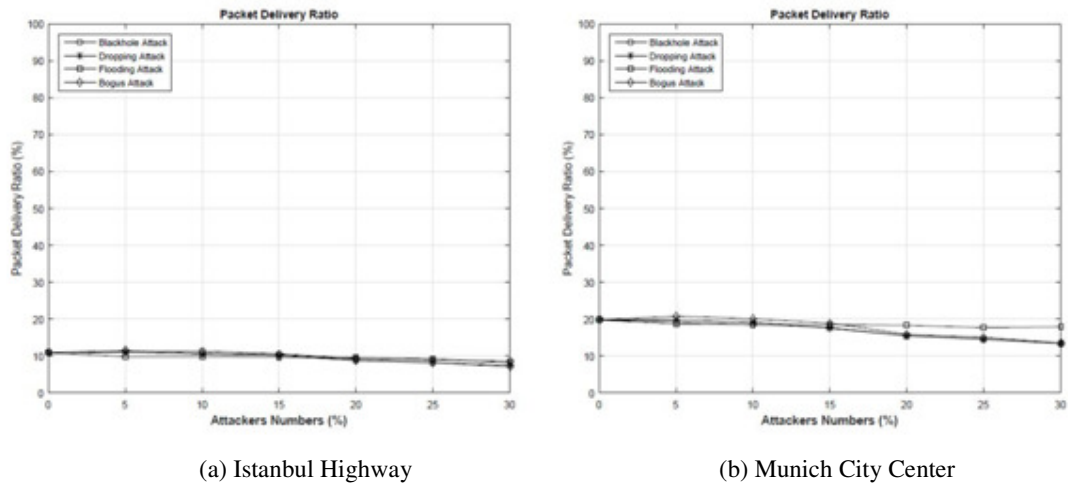


Figure 4. Packet Delivery Ratio – GPSR

Figure 4 shows the packet delivery ratio of all attacks in both maps. GPSR's instantaneous vehicle selection to transmit a packet does not always succeed. Lack of selecting the best route for the destination might result in poor packet delivery performance. As expected the packet delivery ratio was higher on the more dense network. Since a node could find more alternative routes to a destination node in such networks, the sustainability of communication could be extended. More dense networks, consisting of more vehicles, could be more suitable to show the reaction of GPSR against attacks in the future.

GPSR is affected almost equally for all attacks as demonstrated in Figure 4. The main difference between AODV and GPSR is that AODV has a pre-route establishment, where routes are established before the packet transmission begins. For this reason AODV has higher packet delivery ratio than GPSR. Also, the density of networks is significant to the packet delivery ratio. Since a node could find more alternative routes to a destination in dense networks, the sustainability of communication could be provided for longer.

5.3.2 Overhead – GPSR

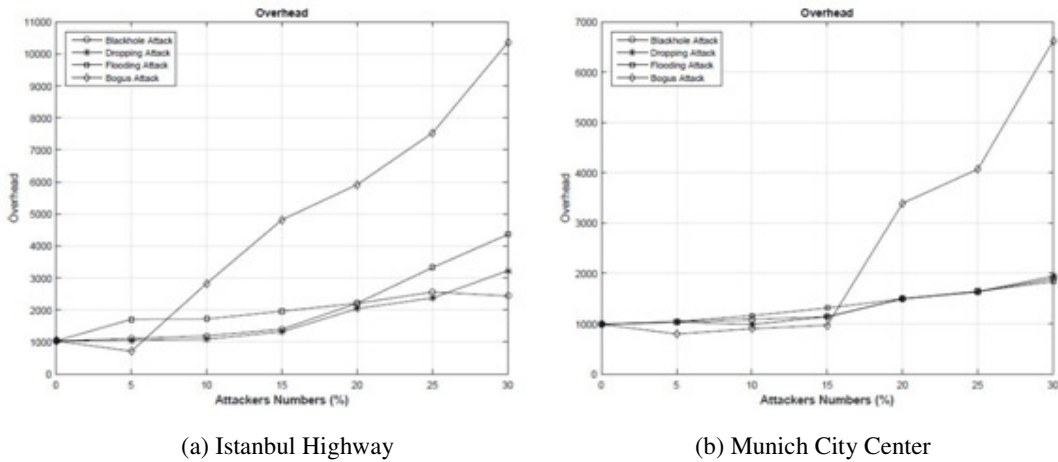


Figure 5. Overhead – GPSR

Overhead results are given in Figure 5 for all attacks in both maps. GPSR clearly has more overhead than AODV. Due to the high number of beacon packets and having two different forwarding mechanisms [25], overhead is quite high in GPSR even when not under attack. When GPSR cannot find a suitable vehicle to transmit a packet, more control packets (beacons) are broadcast to the network. Besides periodic beacon packets, LOCS packets sent more frequently under high mobility is another factor affecting overhead in GPSR. As demonstrated, the overhead of GPSR under attack demonstrates a dramatic increase.

Since there are already more routing control packets in low density networks, they are slightly more affected by flooding attacks in both routing protocols. As the attacker number increases more control packets will be burst to the network, which resulting in increased overhead. Moreover, this attack is more damaging in GPSR as the attacker sends beacon packets to all its neighbors. The increase in the routing control packets can clearly be seen in Figure 5.

5.3.3 End-to-End Delay – GPSR

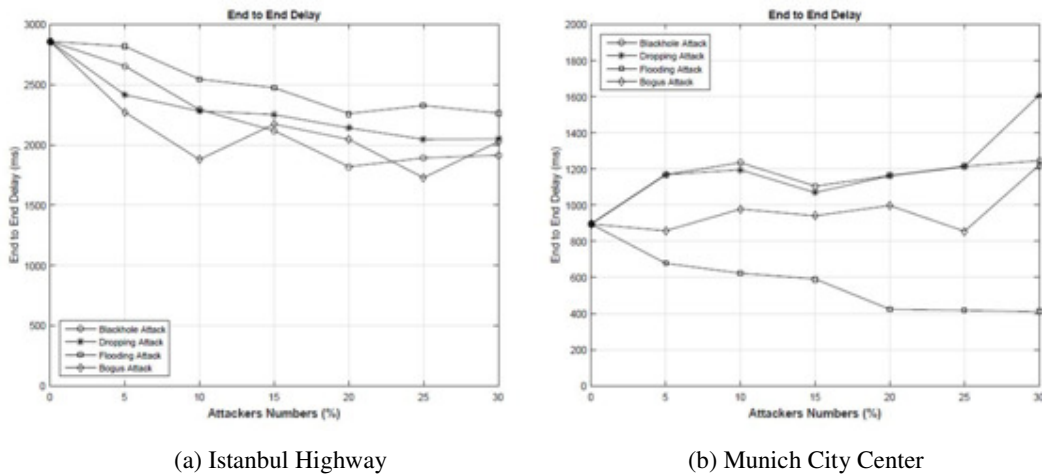


Figure 6. End-to-End Delay – GPSR

Figure 6 shows the end-to-end delay for attacks in the two different maps. In Istanbul Highway map, GPSR's end-to-end delay for all attacks is decreasing. Since fewer packets are transmitted over a short period time to the destination point, end-to-end delay is decreasing. On the other hand, Munich City Center is less not affected than Istanbul Highway due to the high density of nodes in the city center traffic, and more application of GPSR's greedy forwarding mechanism under attack. It should be noted that density is not the only major factor affecting end-to-end delay. There are also other parameters such the location of attackers, the network topology, and traffic patterns.

To summarize up, each attack negatively affects the communication in vehicular ad hoc networks. AODV is generally more severely affected by routing attacks. On the other hand, AODV has a better packet delivery ratio than GPSR in a network under no attack. This is because GPSR does not always select the best route as it decides packet transmission location instantaneously. As expected, results showed that both protocols have better performance in dense networks under no attack. Although AODV demonstrates fairly good performance on networks under no attack, the pre-establishing mechanism of AODV shows a weakness which attackers could exploit. On the other hand, the instantaneous path selection mechanism of GPSR hardens attackers to put themselves in a path. The attacker could directly change the communication links to its neighbors only. In the results, the attack which affects AODV the most is a blackhole attack. In AODV, an attacker has a high chance of diverting the packet transmission by sending fake RREP packets. GPSR are generally affected by each attack, especially when the percentage of attackers in the network exceeds 20% of all nodes. More dense networks consisting of more vehicles could be more suited to showing the reaction of GPSR against attacks.

6. CONCLUSION

Vehicular ad hoc networks are an emerging technology which it is believed will be extensively used in the near future. However, security is a key issue that first needs to be addressed. In order to be able to develop suitable prevention and detection mechanisms for VANETs, the nature of attacks and their effects on the network should be carefully analyzed; and which was the primary aim of this study. The attacks, namely blackhole, dropping, flooding and bogus information, are implemented on AODV and GPSR routing protocols. Although there has been some analyses of attacks specific to MANETs, their effects on more dynamic environments are lacking in the literature, hence they were explored in this current study. More popular attacks against VANETs such as bogus information attacks are also implemented and analyzed. More importantly, all attacks were implemented on real maps and under realistic scenarios. Furthermore, the impacts of the number of attackers and the density of road traffic are shown in the results. Especially GPSR is affected when the number of attackers exceeds 20% of the network. For AODV, the attack type is more influential in such experimental settings. The subtle attacks such as blackhole attack decrease the performance of AODV dramatically. The simulation results clearly show the need of security mechanism suitable for a such highly dynamic environment. To the best of the authors' knowledge, this current study will be one of the most extensive attack analyses for VANETs to be found in the literature, helping future researchers working in this area.

REFERENCES

- [1] R. Engoulou Gilles, M. Bellache, S. Pierre, and A. Quintero, "Vanet Security Surveys," *Computer Communications*, vol. 44, pp. 1 – 13, 2014.
- [2] S. Zeadally, R. Hunt, Y.-S. Chen, A. Irwin, and A. Hassan "Vehicular ad hoc networks (vanets): status, results, and challenges," *Telecommunication Systems*, vol. 50, no 4, pp. 217 – 241, 2010.
- [3] M. Raya and J.-P. Hubaux, "Securing vehicular ad hoc networks," *Journal of Computer Security*, vol. 15, no. 1, pp. 39 – 68, 2007.

- [4] R. D. Pietro, S. Guarino, N. Verde, and J. Domingo-Ferrer, "Security in wireless ad-hoc networks a survey," *Computer Communications*, vol. 51, pp. 1 – 20, 2014.
- [5] P. Ning and K. Sun, "How to misuse aodv, a case study of insider attacks against mobile ad hoc routing protocols," *Ad Hoc Networks*, vol. 3, no. 6, pp. 795 – 819, 2005.
- [6] E. F. Ahmed, R. A. Abouhogail, and A. Yahya, "Performance evaluation of blackhole attack on vanet's routing protocols," *International Journal of Software Engineering and Its Applications*, vol. 8, no. 9, pp. 39 – 54, 2014.
- [7] V. Bibhu, R. Kumar, B. S. Kumar, and D. K. Singh, "Performance analysis of black hole attack in vanet," *International Journal Of Computer Network and Information Security*, vol. 4, no. 11, pp. 47–54, 2012.
- [8] J. Hortelano, J. C. Ruiz, and P. Manzoni, "Evaluating the usefulness of watchdogs for intrusion detection in vanets," in *Proceedings of The Conference on Communications Workshops (ICC)*, IEEE, 2010, pp. 1–5.
- [9] P. Yi, Z. Dai, S. Zhang, and Y. Zhong, "A new routing attack in mobile ad hoc networks," *International Journal of Information Technology*, vol. 11, no. 2, pp. 83–94, 2005.
- [10] M. Abdelshafy and P. King, *Resisting flooding attacks on AODV*. International Academy, Research and Industry Association, IARIA, 2014, pp. 14–19.
- [11] A. Sinha and S. K. Mishra, "Preventing vanet from dos & ddos attack," *International Journal of Engineering Trends and Technology (IJETT)*, vol. 4, no. 10, pp. 4373–4376.
- [12] F. Sakiz and S. Sen, "A survey of attacks and detection mechanisms on intelligent transportation systems: Vanets and IoV," *Ad Hoc Networks*, vol. 61, pp. 33–50, 2017.
- [13] R.W. van der Heijden, S. Dietzel, T. Leinmüller, F. Kargl. "Survey on misbehavior detection in cooperative intelligent transportation systems," *arXiv preprint arXiv:1610.06810* (2016).
- [14] B. T. Sharef, R. A. Alsaqour, and M. Ismail, "Vehicular communication ad hoc routing protocols: A survey," *Journal of Network and Computer Applications*, vol. 40, pp. 363 – 396, 2014.
- [15] C. E. Perkins and E. M. Royer, "Ad-hoc on-demand distance vector routing," in *Proceedings of The 2nd International IEEE Workshop on Mobile Computing Systems and Applications (WMCSA)*. IEEE, 1999, pp. 90–100.
- [16] B. Karp and H.-T. Kung, "Gpsr: Greedy Perimeter stateless routing for wireless networks," in *Proceedings of the 6th annual international conference on Mobile computing and networking (MobiCom)*. ACM, 2000, pp. 243–254.
- [17] H. Ghafoor and K. Aziz, "Position-based and geocast routing protocols in vanets," in *Proceedings of the 7th International Conference on Emerging Technologies (ICET)*. IEEE, 2011, pp. 1–5.
- [18] W. Kieß, H. Fußler, J. Widmer, and M. Mauve, "Hierarchical location service for mobile ad-hoc networks," *ACM SIGMOBILE mobile computing and communications review*, vol. 8, no. 4, pp. 47–58, 2004.
- [19] T. Moore, M. Raya, J. Clulow, P. Papadimitratos, R. Anderson, and J. P. Hubaux, "Fast exclusion of errant devices from vehicular networks," in *Proceedings of the 5th International Conference of Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, 2008, pp. 135–143.
- [20] "The Network Simulator ns-2," <http://www.isi.edu/nsnam/ns/>, 2017.

- [21] P. K. Singh, "Article: Influences of tworayground and nakagami propagation model for the performance of adhoc routing protocol in vanet," *International Journal of Computer Applications*, vol. 45, no. 22, pp. 1–6, 2012.
- [22] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, "Sumo–simulation of urban mobility," in *Proceedings of The 3rd International Conference on Advances in System Simulation (SIMUL)*, 2011.
- [23] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *Pervasive Computing*, vol. 7, no. 4, pp. 12–18, 2008.
- [24] S. Sen, J. A. Clark, and J. E. Tapiador, "Security threats in mobile ad hoc networks," *Security of Self-Organizing Networks: MANET, WSN, WMN, VANET*, Auerbach Publications, pp. 127–147, 2010.
- [25] M. R. Jabbarpour, A. Jalooli, E. Shaghghi, A. Marefat, R. M. Noor, and J. J. Jung, "Analyzing the impacts of velocity and density on intelligent position-based routing protocols," *Journal of Computational Science*, vol. 11, pp. 177 – 184, 2015.

ANALYSIS OF WORMHOLE ATTACK CONFIRMATION SYSTEM DURING EMAIL DUMPING ATTACK

Divya Sai Keerthi T and Pallapa Venkataram

Electrical Communication Engineering Department,
Indian Institute of Science, Bangalore, India

ABSTRACT

The wormhole attack is a severe attack on an application in a Mobile Ad hoc Network (MANET). This attack causes the applications to choose longer routes and disturbs the communications. A wormhole attacker can cause havoc on a MANET even without compromising the host of the application. For a wormhole attacker, email dumping is a simple attack that can lead to disastrous effects. In this paper we demonstrate the working of wormhole attack confirmation system in case of email dumping attack. The proposed method uses the honeypot to keep the attackers busy by interacting with them, and simultaneously identifies the attack using attack tree. It further reduces the false alarms, using the history of past attacks, stored in the Attack History Database. The system was tested in various sizes of MANETs, and the results prove that, the system efficiently identifies the email dumping attack with reduce false alarms.

KEYWORDS

Email Dumping Attack, Wormhole attack, Honeypot, Attack Tree, Attack History, MANET.

1. INTRODUCTION

Emails are the most common ways to exchange information, electronics documents and other files. Now-a-days, emails are used for sharing confidential data to electronic ads. In such cases, the emails attract a lot of attackers. Different types of attacks are possible on emails today[1], such as email dumping attack also called as email bomb attack[2], email malware attack[3], email virus[4], phishing emails[5] etc. The effect of these attacks increase drastically in the wireless networks domain like, Mobile Ad hoc Networks (MANETs).

MANETs are increasing becoming the main tools for network centric warfare [6]. The flexibility provided in the design of MANETs makes it easy to execute attacks [7]. Wormhole attack is one of the most complicated attacks on a MANET. It is a routing manipulation attack that has a capacity to control the routes in a MANET. The wormhole attack is launched by two nodes cooperating with each other in order to execute the attack by forming a channel between themselves [8]. This channel is called the Wormhole tunnel. When a wormhole attacker on one end of the tunnel receives a packet, it forwards the packet to the other attacker via the wormhole tunnel, without following any protocol specifications. Hence the route via the tunnel seems to be shorter or quicker route. With the help of this tunnel, the wormhole attacker attracts more routes via themselves. Once a node chooses the path via the wormhole tunnel, the attackers control the application and data transfer.

Using the privileges acquired while executing the wormhole attack, the attackers launch more complicated attacks like the email dumping attack. An email dumping attack (also called as the email bomb attack) is a form of memory dumping attack, where the victim is overwhelmed with emails, i.e., the victim receives excessive emails [9]. The victim could be a server or a destination node. This causes the victim's incoming email buffer to overflow and lead to delay or loss of new emails. The network may be congested, as the lost emails will be sent again causing further delay.

1.1. Proposed Method

In our previous work, we have proposed the Wormhole Attack Confirmation system, which protects the MANET from wormhole attacks, and also protects the benign nodes from being framed as the wormhole attacker. In this paper we aim to analyse the Wormhole Attack Confirmation system during the email dumping attack. The system aims to confirm the email dumping attack using honeypot. The honeypot analyse various features of an email dumping attack using the wormhole attack trees. It further confirms the email dumping attack using the Attack History Database (AHD). The contributions of the paper are as follows:

1. Detection of relevant symptoms during email dumping attack in a MANET.
2. Construction of attack tree using the symptoms of email dumping attack.
3. Analysis of Wormhole Attack Confirmation System in the presence of email dumping attack.

2. RELATED WORKS

The wormhole attack is one of the most common attacks in a MANET environment. Many authors have published works on methods to identify or detect the wormhole attack, and many survey papers are available to gain an understanding of the landscape of research[10] on wormhole attack. In this paper here, we are interested in discussing about the email dumping attack in particular.

In paper [11], Dwork and Naor have proposed a method to avoid unwanted junk mail from flooding a users' inbox. The proposed method controls the access to common pool of resources by enforcing the user to compute a moderately hard function called the pricing function for important resources, and shortcut function for cheap resources. Various functions such as, extracting the square root, Fiat-Shamir based scheme, Ong-Schnorr-Shamir based scheme and recycling broken signature we tested, of which Fiat-Shamir scheme performed most efficiently.

Jakobsson and Menzer have given a detailed account on how an attacker executes an attack, in which the victim is bombarded with un-wanted mails in [12]. The author explain, how the attacker first finds the suitable forms, which take victim address as input, and how these forms are filled, automatically using scripts. The authors' term this as poor man's DoS and explained how it is different from regular DoS. The paper also suggest some lightweight method to prevent and detect such attacks by, avoiding emails via open relays or using CAPTCHA or using extended address book at the user end.

In [13], Chinchani et.al, have proposed methods to analyse the insider threat, which is generally ignored by many organizations. The working of the proposed scheme is analysed using the example of email worms, a resources-based attack. The KH model proposed in the paper, places a constraints that the attack will be successful when attacker can compromise all the reachable nodes via email. Thus in order to stop this attack the author suggests that mail server randomly

drops mails of at least one victim, i.e., at least one victim remains not reachable, thereby failing the attack.

[14] is a narrative of an email spam attack that took place on Langley AFB internetworking infrastructure. The authors explained in chronological order, the events that happened during the attack; and, how they updated the countermeasure strategy each time, when the attacker improved their strategy. Finally they have designed a filtering algorithm which could mitigate a large variety of email-bomb attacks.

The paper [15] proposes a Progress Email Classifier (PEC) for differentiating between the good emails and unsolicited bulk emails. The proposed classifier maintains a scoreboard for feature instances of email, which are used to classify the mail. The email classified as unsolicited bulk email is passed to a blacklist for further handling.

The author of paper [16], have proposed a method to detect the email spam using data mining and machine learning. Three classifiers were used for identifying email spam, naive bayes, sequential minimal optimization and J48. Out of the three classifiers, J48 performed well compared to the other two methods.

3. PRELIMINARIES

When an attacker launches a new attack by using the privileges gained from an earlier attack, it becomes tricky to identify the new attack. In our work proposed in this paper, we present a method to identify the email dumping attack launched using the wormhole attack. In order to identify the email dumping attack launched using the wormhole attack, we use the Wormhole Attack Confirmation System proposed in our previous work (currently under review).

The Wormhole Attack Confirmation system aims to confirm the wormhole attack using the honeypot. The honeypot interacts with the attacker, mimicking as the victim node, while it confirms the attack. To analyse the current attack scenario, honeypot identifies the symptoms of the wormhole attack using the Wormhole Attack Tree. The following are the symptoms of wormhole attack: (a) S_1 Low hop count route replies, (b) S_2 Increased packet delivery time, (c) S_3 RREQ dropped by malicious node, (d) S_4 increased number of neighbours, (e) S_5 Presence of asymmetrical links, (f) S_6 Longer propagation delays, (g) S_7 Reception of same message, (h) S_8 More load on certain nodes. The honeypot further confirms the attack using the Attack History Database.

4. CONFIRMATION EMAIL DUMPING ATTACK USING WORMHOLE ATTACK CONFIRMATION SYSTEM

In this section we prove the efficiency of the Wormhole Attack Confirmation (WAC) system in identifying the email dumping attack. We identify the symptoms of the email dumping attack, and the corresponding symptoms of the wormhole attack. The symptoms of the email dumping attack are modelled using the Wormhole Attack Tree (WAT) and the symptoms of wormhole attack which cause them. The honeypot calculates the strength of the symptoms of wormhole attack using the Wormhole Attack Confirmation system. In what follows, we discuss the symptoms of the email dumping attack and the corresponding wormhole attack trees.

Table 1. Nomenclature used

| Symbol | Description |
|-----------|---|
| S_i | i^{th} symptom of wormhole attack |
| SE_i | i^{th} symptom of email dumping attack |
| $K(S_i)$ | Strength of i^{th} symptom of wormhole attack |
| $K(SE_i)$ | Strength of i^{th} symptom of email dumping attack |
| $K(ED)$ | Strength of email dumping attack from attack tree analysis |
| μ | Weight assigned to symptoms of wormhole attack |
| α | Severity of attack as seen in attack history database |
| P_{ED} | Overall strength of email dumping attack |

4.1. Identifying the Email Dumping Attack

An email dumping attack is a form of denial of service attack. In this attack, the victim receives excessive emails from different nodes in the MANET. This leads to overflowing of incoming email buffer of the victim, loss of emails or delay of emails etc. The symptoms of the email dumping attack are the effects of the attack seen at the victim node. Different execution of the wormhole attack leads to different symptoms of the email dumping attack. Thus, each of the symptom can be modelled into the underlying wormhole attack, which makes it possible. Table 1 provides the list of symptoms of email dumping attack. Let's discuss each symptom in detail.

Table 2. Symptoms of Email Dumping Attack

| Symptom of Email Dumping Attack | Description |
|---------------------------------|--------------------------------|
| SE_1 | Over flowing buffer |
| SE_2 | Increased email arrival rate |
| SE_3 | Emails from various sources |
| SE_4 | Emails of various destinations |
| SE_5 | Delay in emails |
| SE_6 | Loss of emails |
| SE_7 | Slow network operations |

4.1.1. Over flowing buffer

Due to the excessive number of emails delivered to the victim, the buffer of the victim is usually full in an email dumping attack. This symptom is caused when the victim node recursively receives the same message or when a node handles many routes in the MANET. The strength of over flowing buffer symptom SE_1 , $K(SE_1)$ given as follows:

$$K(SE_1) = K(S_8) + K(S_7) - (K(S_8) * K(S_7))$$

4.1.2. Increased email arrival rate

The main characteristic of an email dumping attacker is to send large number of emails to the victim, at a faster speed. Thus the arrival rate of the emails is high in this attack. This symptom is caused when the rate of arrival of emails increased drastically. The arrival rate of emails increases when: (a) a particular node has shorter distance to other nodes and has many neighbours or (b) a node is receiving multiple copies of the same message due to lack of acknowledgement at the sender. The strength of increased email arrival symptom SE_2 , $K(SE_2)$ given as follows:

$$K(SE_2) = (K(S_1) * K(S_4)) + K(S_7) - (K(S_1) * K(S_4) * K(S_7))$$

4.1.3. Emails from various sources

To overwhelm the victim, the attacker creates route such that mails of various sources get dumped at the victim. Emails from various sources arrive at victim node, when it has more number of neighbours and other nodes in the region are not forwarding the mails. The strength of the symptom SE_3 , $K(SE_3)$ given as follows:

$$K(SE_3) = (K(S_3) * K(S_4))$$

4.1.4. Emails of various destinations

Many nodes choose a particular node as a hop to reach various destinations when, (a) the node has shortest path to the destination and has more neighbours or (b) when it promptly forwards the mail to all its neighbours. The strength of the symptom SE_4 , $K(SE_4)$ given as follows:

$$K(SE_4) = (K(S_1) * K(S_4)) + (K(S_3) * K(S_4)) - ((K(S_1) * K(S_4)) * (K(S_3) * K(S_4)))$$

4.1.5 Delay in emails

A delay in delivery of emails is caused when the network has asymmetrical links; or, has longer propagation delay in some links; or, general packet delivery time is more. The strength of the symptom SE_5 , $K(SE_5)$ given as follows:

$$K(SE_5) = K(S_2) + K(S_6) + K(S_5) - (K(S_2) * K(S_6)) - (K(S_6) * K(S_5)) - (K(S_5) * K(S_2)) + (K(S_2) * K(S_6) * K(S_5))$$

4.1.6. Loss of emails

Emails forwarded to a victim node are lost when nodes drop the messages received by them or when certain nodes handle too many emails, and drop a few in the processing. The strength of the symptom SE_6 , $K(SE_6)$ given as follows:

$$K(SE_6) = K(S_8) + K(S_3) - (K(S_8) * K(S_3))$$

4.1.7. Slow network operations

Network operation, during an email dumping attack, slows down due to the excessive load on the victims in the network. The strength of the symptoms $K(SE_7)$ is given as follows

$$K(SE_7) = K(S_8)$$

The overall strength of the email dumping attack identified by the wormhole attack confirmation system is given by $K(ED)$. The overall strength of email dumping attack, $K(ED)$ is given as:

$$K(ED) = \sum_i K(SE_i) - \sum_{j \leq i} K(SE_i) K(SE_j) + \sum_{k \leq j \leq i} K(SE_i) K(SE_j) K(SE_k) - \dots + \prod_i K(SE_i)$$

4.2. Confirming the Email Dumping Attack

Once the email dumping attack is identified, honeypot confirms the occurrence of the email dumping attack, considering the input from the Attack History Database (AHD). It analyses the current strength of the email dumping attack in the context of previous attacks recorded in the AHD. After analysis honeypot takes a decision on the occurrence of the attack.

Table 3. Classes of Symptom Strength

| Strength of Wormhole Attack Symptoms | Class |
|--------------------------------------|----------|
| (0, 0.3] | Low |
| (0.3, 0.7] | Moderate |
| (0.7, 1] | High |

The honeypot analyses the strength of the email dumping attack, $K(ED)$, with respect to the strength of wormhole attack symptoms (see table 2). According to the history of email dumping attack, the following weights are assigned to the intervals of $K(ED)$.

(a) Weak symptoms of wormhole attack:

$$\mu = \begin{cases} 1, & K(ED) \leq 0.3 \\ \frac{0.5 - K(ED)}{0.2}, & 0.3 < K(ED) < 0.5 \\ 0, & 0.5 \leq K(ED) \leq 1 \end{cases}$$

(b) Moderate strength of wormhole attack symptoms:

$$\mu = \begin{cases} 0, & K(ED) \leq 0.1 \\ \frac{K(ED) - 0.1}{0.2}, & 0.1 < K(ED) < 0.3 \\ 1, & 0.3 \leq K(ED) \leq 0.7 \\ \frac{0.9 - K(ED)}{0.2}, & 0.7 < K(ED) < 0.9 \\ 0, & 0.9 \leq K(ED) \end{cases}$$

(c) High strength of wormhole attack symptoms:

$$\mu = \begin{cases} 0, & K(ED) \leq 0.5 \\ \frac{K(ED) - 0.5}{0.2}, & 0.5 < K(ED) \leq 0.7 \\ 1, & 0.7 < K(ED) \end{cases}$$

The honeypot then queries the AHD for similar attacks in the past. The inputs from the AHD are analysed for any past attacks on MANET. A severity value, α , is assigned to the attacks that occurred in the past, as shown in Table 3.

Table 4. Similar attacks in history

| Number of attacks in the past $N(\text{attack})$ | Severity α |
|---|----------------------|
| $0 < N(\text{attack}) < 3$ | 0.3 |
| $3 \leq N(\text{attack}) < 10$ | 0.7 |
| $10 \leq N(\text{attack})$ | 1 |

Finally the overall strength of the email dumping attack, P_{ED} is given as:

$$P_{ED} = \mu * K(ED) + (1 - \mu) * (\alpha)$$

If the $P_{ED} > 0.7$, the email dumping attack is confirmed and Honeypot starts to trace the location of the attacker, which is a future work. If P_{ED} is in between $[0.3, 0.7)$ then it is considered as weak confirmation and the Honeypot continues to interact with the attacker to improve the information about the attacker. If $P_{ED} < 0.3$ then Honeypot discards the observations as false alarms.

5. SIMULATION RESULTS

5.1. Simulation Scenario

The simulation scenario consists of MANET with size varying from 50 to 200 nodes and each node's speed varies from 10m/sec to 20m/sec. A node in the MANET holds an email server. Two nodes at random are chosen to act as wormhole attackers executing the email dumping attack. A resource rich node in the centre of the MANET is chosen as the honeypot. Table 4 lists the remaining parameters of the simulation scenario.

Table 5. Parameters of Simulation

| Parameter | Value |
|-------------------------------|------------------|
| Number of nodes | 50-200 |
| MAC protocol | 802.11a |
| Routing protocol | AODV |
| Traffic source | CBR |
| Path-loss model | Two-ray |
| Mobility model | Random way point |
| Radio Range | 270m-300m |
| Packet size | 512 bytes |
| Speed | 10,15,20 m/s |
| Queuing Policy at the routers | FIFO |
| Channel capacity | 2Mbits/s |

5.2. Simulation Results

Figure 1 shows the percentage of email dumping attacks confirmed using the Wormhole Attack Confirmation system. The system confirms all the email dumping attacks with a minimum of 4 symptoms of wormhole attack. This shows the ability of the system to confirm the attack quickly and efficiently.

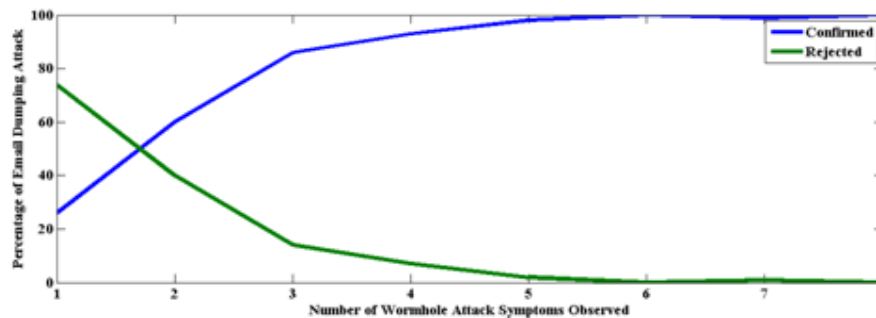


Figure 1. Percentage of Email Dumping Attack Accepted and Rejected

The number of symptoms of email dumping attack, identified with symptoms of wormhole attack is shown in figure 2. With just one symptom observed, the system can identify around 3 symptom of email dumping attack. The best case performance of the system is when all the symptoms of

email dumping attack are can be identified with just 5 symptoms of the wormhole attack. This shows that the model presented in the paper is efficient at deducing the email dumping attack

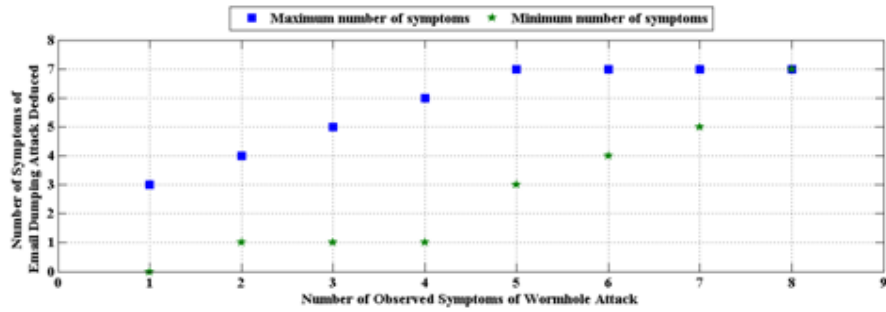


Figure 2. Number of Symptoms of Email Dumping Attack Identified

Figures 3, 4 and 5 show the effect of history on the confirmation of email dumping attack. In order to confirm the attack, in case of weak and moderately strong email dumping attack symptoms, the system needs at least 3 symptoms to confirm the email dumping attack. However, the best case performance of the system is achieved when strong email dumping attack symptoms are available in the history. When the symptoms are strong, the email dumping attack can confirmed even when the attack was observed just once in the past.

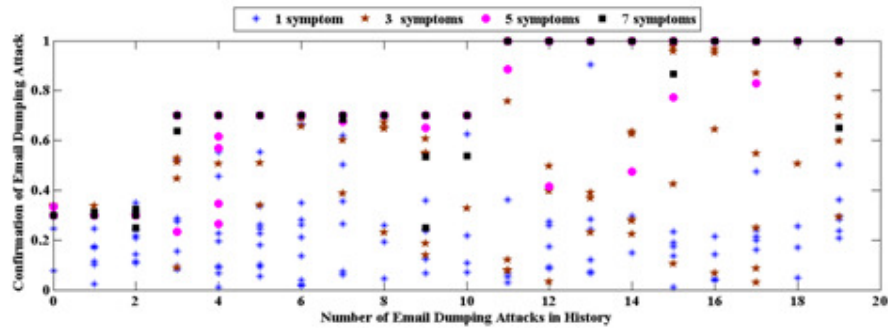


Figure 3. Effect of Attack History in presence of Weak Symptoms

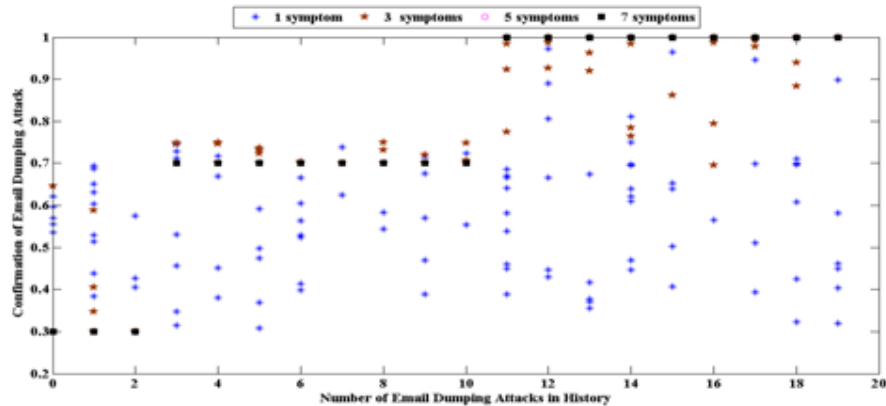


Figure 4. Effect of Attack History in presence of Moderate Symptoms

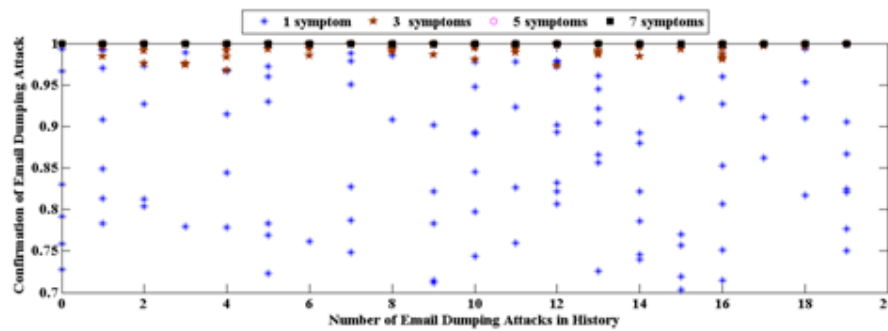


Figure 5. Effect of Attack History in presence of Strong Symptoms

6. CONCLUSIONS

The study presented in this paper gives a detailed analysis of the Wormhole Attack Confirmation system during the email dumping attack. Various scenarios of the email dumping attack were modelled using the symptoms of the wormhole attack. The results presented in the paper show that the system is capable of confirming the email dumping attack with a high probability, in most of the cases. This shows that the Wormhole Attack Confirmation system is capable of identifying and confirming the email dumping attack.

REFERENCES

- [1] Shaun Aimoto, Tareq AlKhatib, Peter Coogan, Mayee Corpin, Jon DiMaggio, Stephen Doherty, Tommy Dong, James Duff, Brian Fletcher, Kevin Gossett, Sara Groves, Kevin Haley, et al. "Symantec internet security threat report trends for 2017." Volume XXII (2017).
- [2] Massive Email Bombs Target .Gov Addresses, 2016, Online: <http://krebsonsecurity.com/2016/08/massiveemail-bombs-target-gov-addresses/>
- [3] Wen, Sheng, Wei Zhou, Jun Zhang, Yang Xiang, Wanlei Zhou, Weijia Jia, and Cliff C. Zou. "Modeling and analysis on the propagation dynamics of modern email malware." *IEEE transactions on dependable and secure computing* 11, no. 4 (2014): 361-374.
- [4] Zou, Cliff C., Don Towsley, and Weibo Gong. "Email virus propagation modeling and analysis." Department of Electrical and Computer Engineering, Univ. Massachusetts, Amherst, Technical Report: TR-CSE-03-04 (2003).
- [5] Qabajeh, Issa, and Fadi Thabtah. "An experimental study for assessing email classification attributes using feature selection methods." In *Advanced Computer Science Applications and Technologies (ACSAT), 2014 3rd International Conference on*, pp. 125-132. IEEE, 2014.
- [6] Shams, Tariq Ali, and Adnan K. Kiani. "Routing over intermittent links for network centric warfare applications." In *Wireless Communications and Networking Conference (WCNC), 2014 IEEE*, pp. 2224-2229. IEEE, 2014.
- [7] Keerthi, T. Divya Sai, and Pallapa Venkataram. "AODV route maintenance using HoneyPots in MANETs." In *Internet Security (WorldCIS), 2015 World Congress on*, pp. 105-112. IEEE, 2015.
- [8] Keerthi, T. Divya Sai, and Pallapa Venkataram. "Locating the attacker of wormhole attack by using the honeypot." In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on*, pp. 1175-1180. IEEE, 2012.

- [9] Aljumah, Abdullah, and Tariq Ahamad. "A Novel Approach for Detecting DDoS using Artificial Neural Networks." *International Journal of Computer Science and Network Security* 16, no. 12 (2016): 132-138.
- [10] Shrivastava, Akansha, and Rajni Dubey. "Wormhole Attack in Mobile Ad-hoc Network: A Survey." *International Journal of Security and Its Applications* 9, no. 7 (2015): 293-298.
- [11] Dwork, Cynthia, and Moni Naor. "Pricing via processing or combatting junk mail." In *Annual International Cryptology Conference*, pp. 139-147. Springer, Berlin, Heidelberg, 1992.
- [12] Jakobsson, Markus, and Filippo Menczer. "Untraceable email cluster bombs: On agent-based distributed denial of service." *arXiv preprint cs/0305042* (2003)
- [13] Chinchani, Ramkumar, Duc Ha, Anusha Iyer, Hung Q. Ngo, and Shambhu Upadhyaya. "Insider threat assessment: Model, analysis and tool." In *Network Security*, pp. 143-174. Springer US, 2010.
- [14] Bass, Tim, and Gelln Watt. "A simple framework for filtering queued SMTP mail (cyberwar countermeasures)." In *MILCOM 97 proceedings*, vol. 3, pp. 1140-1144. IEEE, 1997.
- [15] Lin, Sheng-Ya, Cheng-Chung Tan, Jyh-Charn Liu, and Michael Oehler. "High-speed detection of unsolicited bulk emails." In *Proceedings of the 3rd ACM/IEEE Symposium on Architecture for networking and communications systems*, pp. 175-184. ACM, 2007.
- [16] ZhiWei, Mi, Manmeet Mahinderjit Singh, and Zarul Fitri Zaaba. "Email Spam Detection: A Method Of Metaclassifiers Stacking." In *The 6th International Conference on Computing and Informatics*, pp. 750-757. Kuala Lumpur Malaysia, 2017.

AUTHORS

T. Divya Sai Keerthi received her Bachelor of Technology degree from the Jawaharlal Nehru Technological University, Hyderabad, India in 2009. She is currently pursuing her Ph.D in Indian Institute of Science, Bangalore, India. Her research interest include the fields of Wireless and Ad hoc Communication, Communication Protocols, Computation Intelligence applications in Communication Networks and Mathematical Modelling.



Pallapa Venkataram received his Ph.D. Degree in Information Sciences from the University of Sheffield, England, in 1986. He is currently the chairman for centre for continuing education, and also a Professor in the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India. Dr.Pallapa's research interests are in the areas of Wireless Ubiquitous Networks, Social Networks, Communication Protocols, Computation Intelligence applications in Communication Networks and Multimedia Systems. He is the holder of a Distinguished Visitor Diploma from the Orrego University, Trujillo, PERU. He has published over 150 papers in International/national Journals/conferences.



EXTRACTION AND REFINEMENT OF FINGERPRINT ORIENTATION FIELD

Pierluigi Maponi, Riccardo Piergallini and Filippo Santarelli

Department of Mathematics, University of Camerino, Camerino, Italy

ABSTRACT

We propose a gradient-based method to extract the orientation field of a fingerprint image, and an iterative algorithm to refine and regularise this field. The formulation of this iterative algorithm is based on two new integral operators, which are described together with their main properties. A preprocessing step is also proposed in order to enhance the performance of the whole procedure. The results of our tests on real fingerprint images are provided to show the performance of the proposed approach.

KEYWORDS

Fingerprint analysis, Preprocessing, Equalisation, Segmentation, Orientation Extraction, Orientation Refinement

1. INTRODUCTION

Biometrics is a widely studied field, engaging scientists in several disciplines such as engineering, biology, mathematics and law. Interest in biometric authentication or recognition has grown in the last decades for medical, forensic, government and socio-economic applications. Fingerprint is one of the most distinctive biometric trait, different even for twins, with low storage needs and low cost acquisition systems; hence it is one of the most used biometric measure. Recent advances in computer science and improvements of hardware performances allowed the development of several automated fingerprint recognition systems. Such systems operate in two different modes: verification and identification. Verification is mostly used for civilian purposes, such as restricted resources access control, where an input fingerprint is compared with a database of already enrolled fingerprints; in the event that the input is present in the database, the verification is successful and the access granted. Identification is intended to find the identity of a person, and a fingerprint is compared with a database, that may not contain it; this is mainly used by law enforcement agencies for investigation purposes and the analysis of the crime scene.

Fingerprint images are characterised by a very particular structure formed by several almost parallel curves, which are usually called ridges. A typical recognition system consists of the following steps:

1. Image acquisition, through inked paper scanning or directly by newer fingerprint scanners.
2. Preprocessing, aimed to noise removal and contrast enhancement.
3. Feature extraction, where some selected features are computed from the fingerprint for later use in the matching stage, or possibly stored in a database. In this stage is often present a codification procedure to reduce storage needs and to boost matching speed.
4. Postprocessing, intended to improve the extracted features or to remove bad ones.

5. Matching, where a comparison is performed between the extracted features and the ones stored in the database [1-3].

The orientation field estimation, that is the extraction of information about ridge flow, is frequently present inside a fingerprint processing algorithm. Indeed the orientation field is used for classification [4], to detect singular points [5], to detect fingerprints alterations [6], for registration before matching [7], to improve matching performance [8, 9] and as a matching feature in itself [10]. Thus, the computation of a reliable orientation field from a fingerprint image is a problem of great interest and many techniques have been already developed. Usual approaches to this problem are: gradient-based, slit- and projection-based techniques, frequency domain orientation estimation. Poor quality fingerprints may contain creases, scratches, discontinuous ridge patterns and no-signal areas, which yield noise in the orientation fields, regardless of the chosen approach [3]. Hence, several methods have been proposed for post-processing in order to reduce the orientation field unreliability, such as the orientation regularisation by using coherence criteria [11], neural network classification of unreliable orientations [12], multi-scale analysis for the correction of elements that change among different scales [13]. Other interesting contributions come from [14], where the correction of the estimated orientation field is based on the information coming from a global orientation model, and [15], where a probabilistic approach to the orientation field regularisation is introduced. [16] and [3] provide a detailed description of other interesting approach for the orientation field enhancement.

Throughout this paper we focus on the second and third stages of the typical fingerprint recognition system. A new algorithm for the estimation of the orientation field and its refinement is proposed. The estimation of the orientation field is performed by a gradient based method: a group of directional gradient masks are chosen and convolved with the image, and their responses are combined to get an orientation field. The enhancement step is based on two operators that can be applied to the orientation field to detect singular points, to smooth the integral curves around singular points, and to iteratively remove noise.

In Section 2 a detailed description of the proposed method is provided. In Section 3 we describe some numerical results obtained with the proposed method. In Section 4 we provide some conclusions.

2. ALGORITHM

Let I be an $N \times M$ grey scale image, where each element $I(i, j)$ with $i = 1, \dots, N$ and $j = 1, \dots, M$ is a grey level ranging from 0 to 255.

Fingerprint images require the definition of the direction and orientation fields; here we use the same distinction introduced by Sherlock and Monroe [17]. Given a complex number $z \in \mathbb{C}$, its phase is the angle $\theta \in [0, 2\pi)$ that it forms with the positive real axis. Every complex number z having the same phase angle θ defines the same direction, regardless its modulus. The set of all the possible directions can be naturally identified with the unit circle S^1 . Consider the straight line γ given by tz , with $t \in \mathbb{R}$; it forms an angle ϕ with the positive real axis that lies in the range $[0, \pi)$, since γ is invariant by rotation through integer multiples of π . Every complex number z having the same angle ϕ defines an orientation, regardless its modulus. The set of all the possible orientations can be naturally identified with the projective circle P^1 . Note that the concept of vector field, that is a mapping from the image to S^1 , is unsuitable to describe the ridge flow of a fingerprint; hence our aim is to compute an orientation field, that is a map from the image to P^1 .

The proposed algorithm is composed by the two steps: preprocessing, described in Section 2.1, and orientation extraction, outlined in Section 2.2.

2.1. Preprocessing

The preprocessing stage combines the image equalisation, the segmentation and the ridges amplification, and is an important preliminary step that yields reliability to the orientation estimation.

The very first operation to be performed is the image scaling to the full grey level range, i.e. 1, ..., 255, applying a linear transformation. The grey level histogram is then computed and smoothed convolving it with a Gaussian kernel.

Fingerprint images coming from real test cases often produce an unbalanced histogram, implying an image either too bright or too dark. So a more balanced distribution of grey levels can be obtained through an iterative process, where an adaptive threshold grey level \bar{x} on the histogram is computed in such a way that \bar{x} is the mean value between the right and the left means; let us call m_L and m_R the final left and right means respectively.

Since small grey values variations around 0 or 255 do not provide very useful information, we can remove them. Given two real numbers $t_L, t_R \in [0,1]$ we compute two threshold values x_L and x_R on the histogram so that $x_L = t_L m_L$ and $x_R = 255 - t_R(255 - m_R)$; we define the following continuous function:

$$f(x) = \begin{cases} 0 & x \leq x_L, \\ \frac{x - x_L}{\bar{x} - x_L} \cdot 127 & x_L < x \leq \bar{x}, \\ 127 + \frac{x - \bar{x}}{x_R - \bar{x}} \cdot 128 & \bar{x} < x \leq x_R, \\ 255 & x > x_R, \end{cases} \quad (1)$$

that maps grey levels less than x_L to 0, \bar{x} to 127, and the ones greater than x_R to 255. Finally we can obtain the equalised image I_E by computing

$$I_E(i, j) = f(I(i, j)), i = 1, \dots, N, j = 1, \dots, M. \quad (2)$$

The preliminary step to compute a significant mask is to remove marginal lines with very small variations in grey values. A minimum-allowed variation threshold τ_V is set; from the image borders to the centre, for each row or column its maximum variation V is computed and if $V < \tau_V$ the row or column is removed.

Acquisitions coming from fingerprint cards, because of camera misalignment, often present oblique lines nearby the border that must be excluded from further analysis. So, we apply to the image a two-dimensional $5 \times \left\lceil \frac{M}{4} + 1 \right\rceil$ filter, where $\lceil \cdot \rceil$ is the rounding operator, with the kernel of the form:

$$(1, 1, 0, -1, -1)^T \cdot (1, 1, \dots, 1, 1), \quad (3)$$

and T is the matrix transposition operator. This filter's response has a high absolute value where there are strong vertical changes in grey values. The outputs of this filter and of the transposed filter are combined together with a line-fitting algorithm to detect oblique lines; the image part outside those lines is removed from further steps.

Fingerprints may contain handwritten text and other artifacts, that are removed in this step. We perform a simple image binarisation and morphology operations to exclude small or thin connected components. The image gradient is also binarised, dilated, and worked with classical morphology techniques, so that we obtain a second mask stricter than the previous. In this way artifacts can be removed and foreground be segmented.

In the following we proceed to amplify the ridges. Let I_R be the restriction of the initial image I to the significant mask obtained in the segmentation step. A matrix I_M is computed by taking the

maximum in a circular neighbourhood of every pixel of I_R ; let I_m be the matrix computed analogously by considering the circular minimum filter applied to I_R . In order to emphasise the alternation of ridges and valleys, two threshold values t_{M_1} and t_{M_2} are chosen for I_M and another two values t_{m_1} and t_{m_2} for I_m ; the values of I_M in the range $[t_{M_1}, t_{M_2}]$ and the values of I_m in the range $[t_{m_1}, t_{m_2}]$ are stretched to the range $[0, 255]$ with a linear scaling. So, the emphasised image I_E can be computed as:

$$I_E(i, j) = I_{m,2}(i, j) + \frac{I_R(i, j) - I_m(i, j)}{I_M(i, j) - I_m(i, j)} (I_{M,2}(i, j) - I_{m,2}(i, j)). \quad (4)$$

2.2. Orientation Extraction

Let I be the initial image restricted to the significant mask obtained in the segmentation and with the ridge-valley structure emphasised as previously described.

The orientation extraction procedure is composed of three steps: orientation estimation, spatial period computation, orientation refinement.

2.1.1. Orientation Estimation

We use a directional gradient based approach, where the image convolution is performed by a Gaussian kernel in a given direction, and a Gaussian derivative in the orthogonal direction. Given $r \in \mathbb{R}_+$, we define the function $K_{r,0}: [-r, r]^2 \rightarrow \mathbb{R}$:

$$K_{r,0}(s, t) = d(s) \cdot e^{-\frac{|d(s)|^{2\alpha_1}}{\sigma_1}} \cdot e^{-\frac{|d(t)|^{2\alpha_2}}{\sigma_2}}, \quad (5)$$

where $d(t) = t - \left(\frac{r}{2} + 1\right)$, and σ_1 , α_1 , σ_2 and α_2 are real positive parameters. We select N_A equally spaced angles $\theta_1, \dots, \theta_{N_A} \in [0, \pi)$ and define the following group of directional gradient kernels:

$$K_{r,k}(i, j) = K_{r,0}(i \cos \theta_k + j \sin \theta_k, -i \sin \theta_k + j \cos \theta_k), \quad (6)$$

where $k = 1, \dots, N_A$, $i = 1, \dots, r_1$, $j = 1, \dots, r_2$. The response to k -th kernel gives the directional image derivatives along the direction with angle θ_k .

An orientation estimation is extracted as follows: we convolve the image with each kernel $K_{r,k}$, compute the absolute value of the response, and smooth it with a Gaussian filter; we call W_k the resulting matrix. Let $\phi_k = \left(\theta_k + \frac{\pi}{2}\right) \bmod \pi$; $O_k = W_k e^{2i\phi_k}$ is a complex matrix with high-magnitude elements where ridges flow along the orientation with angle θ_k . The desired orientation field O is computed as:

$$O(i, j) = \frac{\sum_{k=1}^{N_A} W_k(i, j) \cdot e^{2i\phi_k}}{\sum_{k=1}^{N_A} W_k(i, j)} = \frac{\sum_{k=1}^{N_A} O_k(i, j)}{\sum_{k=1}^{N_A} W_k(i, j)}. \quad (7)$$

Notice that orientation angles are doubled in this process, thus giving a continuous field. A final Gaussian smoothing and an absolute value normalisation are performed, then the phase angles of O are halved.

2.1.2. Spatial Period Computation

From the orientation field O we can estimate the distance between two consecutive ridges. Consider the set of equally spaced gridded points on O

$$\{(i_n, j_n) | n = 1, \dots, N_P\}. \quad (8)$$

For each point (i_n, j_n) consider the fixed-length segment centred at (i_n, j_n) , orthogonal to the orientation $O(i_n, j_n)$, and pick N_S points on it. For each of these points an orientation $o_{n,k} \in \mathbb{C}$ and

a value $v_{n,k} \in [0,255]$ with $k = 1, \dots, N_S$ are obtained by respectively interpolating the field \mathcal{O} and the image I .

In order to consider the n -th segment sufficiently reliable, a minimum absolute value threshold \tilde{t} is chosen and the condition $\min\{|o_{n,k}|\}_{k=1,\dots,N_S} > \tilde{t}$ must be fulfilled, otherwise the segment is skipped. For each reliable segment, say the n -th one, consider the discrete signal

$$v_n[k] = v_{n,k} \quad k = 1, \dots, N_S, \quad (9)$$

apply a lowpass filter, and compute the Fourier spectrum. The spatial frequency $f_{s,n}$ along the n -th segment is the first peak after the zero-frequency one. The spatial frequency for the whole image can be computed as

$$f_s = \frac{1}{N_p} \sum_{n=1}^{N_p} f_{s,n}. \quad (10)$$

The spatial period T_s , i.e. the distance between two consecutive ridges, is given by $T_s = \frac{N_S}{f_s}$.

2.1.3. Orientation Refinement

Let F be a rectangular region in the image I , and $\mathcal{F}: F \rightarrow \mathbb{C}$ be an orientation field, defined in F , possibly given by the initial estimation \mathcal{O} ; we denote with $\mathcal{F}(\mathbf{x})$ the orientation $\mathcal{F}(x, y)$ at point $\mathbf{x} = (x, y)^T \in F$.

The refinement process is defined by two operators. Let $R \in \mathbb{N}$, we select N_C points $\mathcal{C}_R = \{\mathbf{r}_i\}_{i=1,\dots,N_C} \subset \mathbb{R}^2$ from the circumference of radius R centred at $\mathbf{0}$. We define the adjuster \mathcal{G}_A of the field $\mathcal{F}(\mathbf{x})$ as the following orientation field:

$$\mathcal{G}_A(\mathbf{x}) = \frac{1}{N_C} \sum_{k=1}^{N_C} \text{sgn}[a(\mathbf{x}, \mathbf{r}_k)] a(\mathbf{x}, \mathbf{r}_k)^2 \mathcal{F}(\mathbf{x} + \mathbf{r}_k), \quad (11)$$

Where

$$a(\mathbf{x}, \mathbf{r}_k) = \Re \left\{ \frac{\mathcal{F}(\mathbf{x} + \mathbf{r}_k) (\mathbf{r}_k \cdot \mathbf{1} - i \mathbf{r}_k \cdot \cdot)}{|\mathcal{F}(\mathbf{x} + \mathbf{r}_k)| \|\mathbf{r}_k\|} \right\}, \quad (12)$$

and $\mathbf{1}$ and \mathbf{i} are the usual vectors of the canonical base for \mathbb{R}^2 , i is the imaginary unit, \Re gives the real part of a complex number, \cdot denotes the inner product, $\|\cdot\|$ is the Euclidean norm in \mathbb{R}^2 and $|\cdot|$ is the absolute value in \mathbb{C} . We call adjusted field the orientation field \mathcal{AF} obtained as:

$$\begin{aligned} \mathcal{AF}_0(\mathbf{x})^2 &= (1 - s) \mathcal{F}(\mathbf{x})^2 + s \mathcal{G}_A(\mathbf{x})^2, \\ \mathcal{AF}(\mathbf{x}) &= \frac{\mathcal{AF}_0(\mathbf{x})}{|\mathcal{AF}_0(\mathbf{x})|} \max(|\mathcal{F}(\mathbf{x})|, |\mathcal{G}_A(\mathbf{x})|), \end{aligned} \quad (13)$$

where $s \in (0,1)$ is a small parameter.

The smoother \mathcal{G}_S is the other operator and it is defined as follows:

$$\mathcal{G}_S(\mathbf{x}) = \frac{1}{N_C} \sum_{k=1}^{N_C} \text{sgn}[f_d(\mathbf{x}, \mathbf{r}_k)] a(\mathbf{x}, \mathbf{r}_k)^2 \mathcal{F}(\mathbf{x} + \mathbf{r}_k), \quad (14)$$

where $a(\mathbf{x}, \mathbf{r}_k)$ is defined as above and

$$f_d(\mathbf{x}, \mathbf{r}_k) = \Re\{\mathcal{F}(\mathbf{x} + \mathbf{r}_k) \overline{\mathcal{F}(\mathbf{x})}\}, \quad (15)$$

where $\overline{\mathcal{F}(\mathbf{x})}$ is the complex conjugate of $\mathcal{F}(\mathbf{x})$. We call smoothed field the orientation field $\mathcal{S}\mathcal{F}$ obtained as:

$$\begin{aligned}\mathcal{S}\mathcal{F}_0(\mathbf{x})^2 &= (1-s)\mathcal{F}(\mathbf{x})^2 + s\mathcal{G}_S(\mathbf{x})^2, \\ \mathcal{S}\mathcal{F}(\mathbf{x}) &= \frac{\mathcal{S}\mathcal{F}_0(\mathbf{x})}{|\mathcal{S}\mathcal{F}_0(\mathbf{x})|} \max(|\mathcal{F}(\mathbf{x})|, |\mathcal{G}_S(\mathbf{x})|).\end{aligned}\quad (16)$$

The key operator of our procedure for orientation refinement is the smoother, that succeeds in reconstructing and giving global coherence to a noisy orientation field. The drawback of its application is the shifting effect it has on loops; since we iteratively apply the smoothing operator, we need a reliable mask where loops are in the background. The adjuster has the converse effect on loops, giving them back their initial position and enhancing their rounded shape.

The adjusting operator has a nice effect on loops, whereas strongly damages the deltas structure, hence it can be used to detect them. Let \mathcal{O}_S be the orientation estimated with directional gradient filters of size $r = \frac{1}{2}T_s$, and \mathcal{O}_B be the orientation estimated with directional gradient filters with size $r = \frac{3}{2}T_s$, where T_s is the spatial period. We compute a difference mask D between the adjusted field $\mathcal{A}\mathcal{O}_S$ of \mathcal{O}_S and the adjusted field $\mathcal{A}\mathcal{O}_B$ of \mathcal{O}_B . Since small variations of the orientation may also appear outside deltas, small connected components are filtered out; a final dilation is performed on the mask, to be sure to cover deltas.

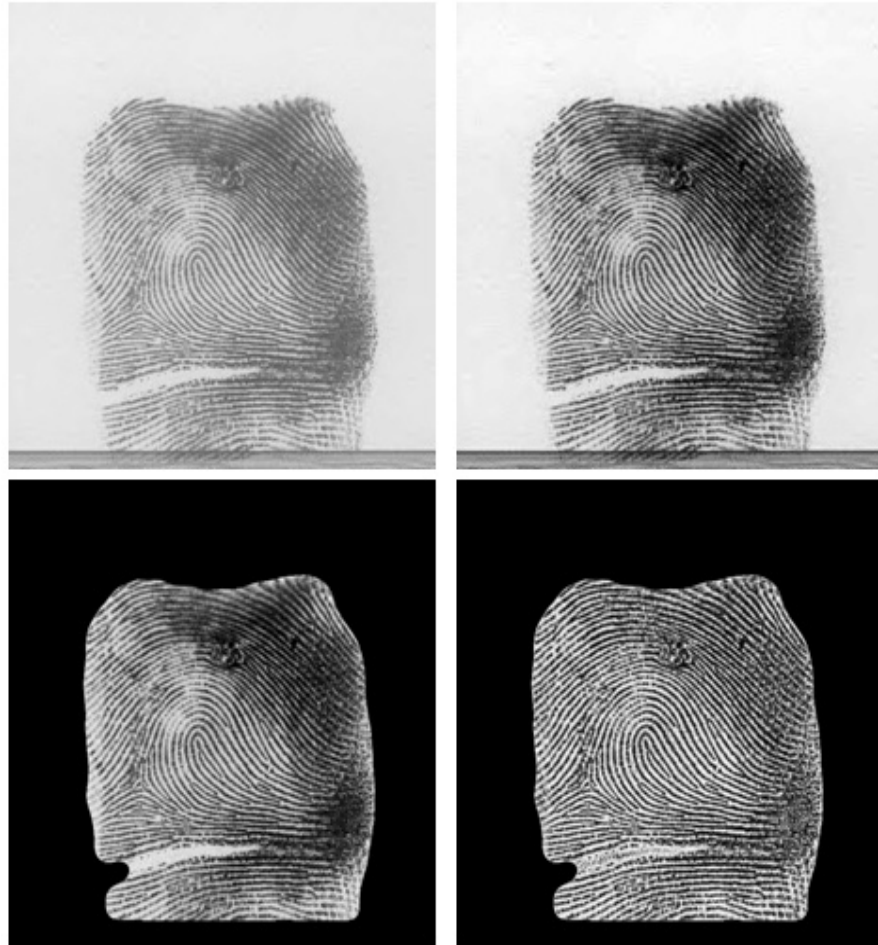


Figure 1. From top left: initial image, equalisation, segmentation and ridge amplification

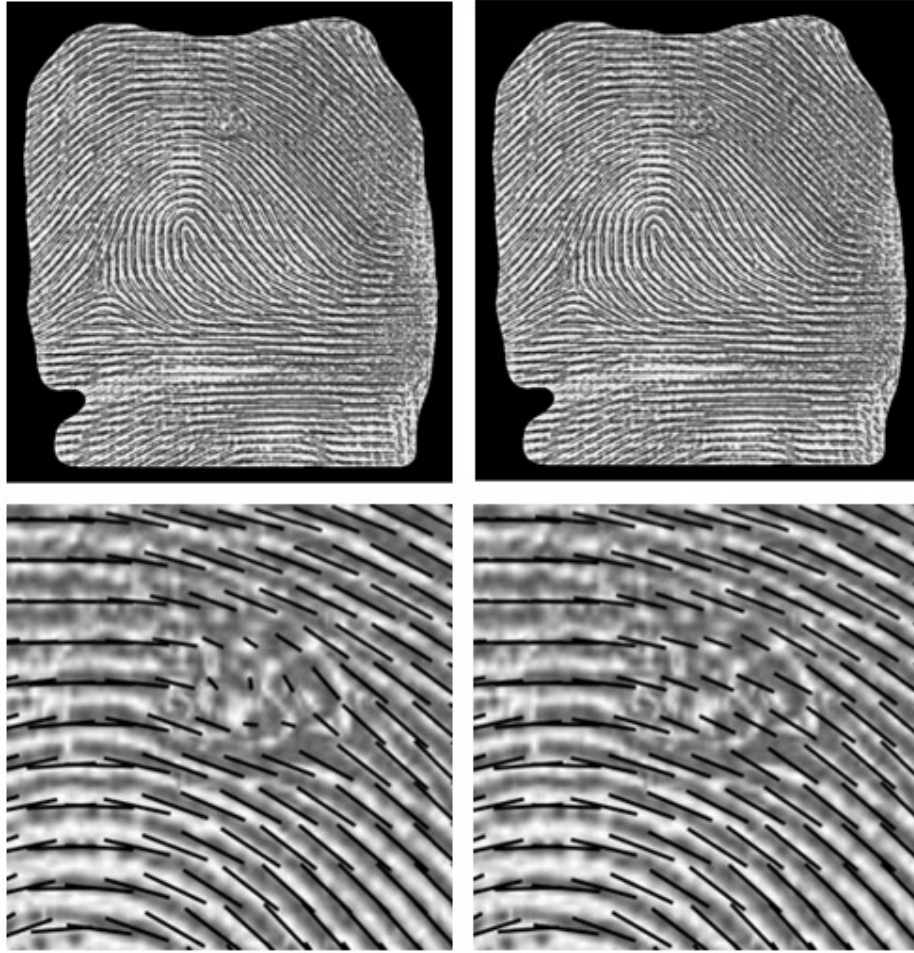


Figure 2. From top left: the extracted orientation field, the result of the refinement procedure, the magnification of a noisy area of the extracted orientation field, and the magnification of the same area after the refinement procedure

A mask M_{CD} with both cores and deltas is computed; M_{CD} is then eroded, so that it does not completely cover singularities. Small connected components, generated by noise and artifacts, are removed.

In our algorithm masks are matrices with values 0 or 255; in the following we use logical operators between masks, thus considering them with false elements in correspondence of 0 values, and with true elements in place of 255 values.

To get a loops-only mask M_L , we need to join the information coming from D , with deltas, and from M_{CD} , with cores too:

$$M_L(i, j) = M_{CD}(i, j) \wedge \text{not}D(i, j), \quad (17)$$

where \wedge and “not” are the usual logical operators. Since during the creation of M_{CD} we eroded the mask, M_L may not entirely cover the loops: we need to perform a dilation. To finally get a no-loops mask M_{NL} we do

$$M_{NL}(i, j) = \text{not}M_L(i, j). \quad (18)$$

An iterative application of the smoothing operator within the mask M_{NL} can deeply improve the orientation estimation. Let us call \mathcal{O}_0 the orientation field obtained in the extraction procedure, and M_0 the starting mask M_{NL} . At the step $k = 1, 2, \dots$ the orientation field \mathcal{O}_k is computed applying the smoothing operator. A difference mask M_k between \mathcal{O}_k and \mathcal{O}_{k-1} is computed and intersected with the previous mask M_{k-1} , then a slight erosion is performed to guarantee convergence. The iterative procedure halts when no more points change orientation due to the smoothing operator, i.e. when M_k is false everywhere.

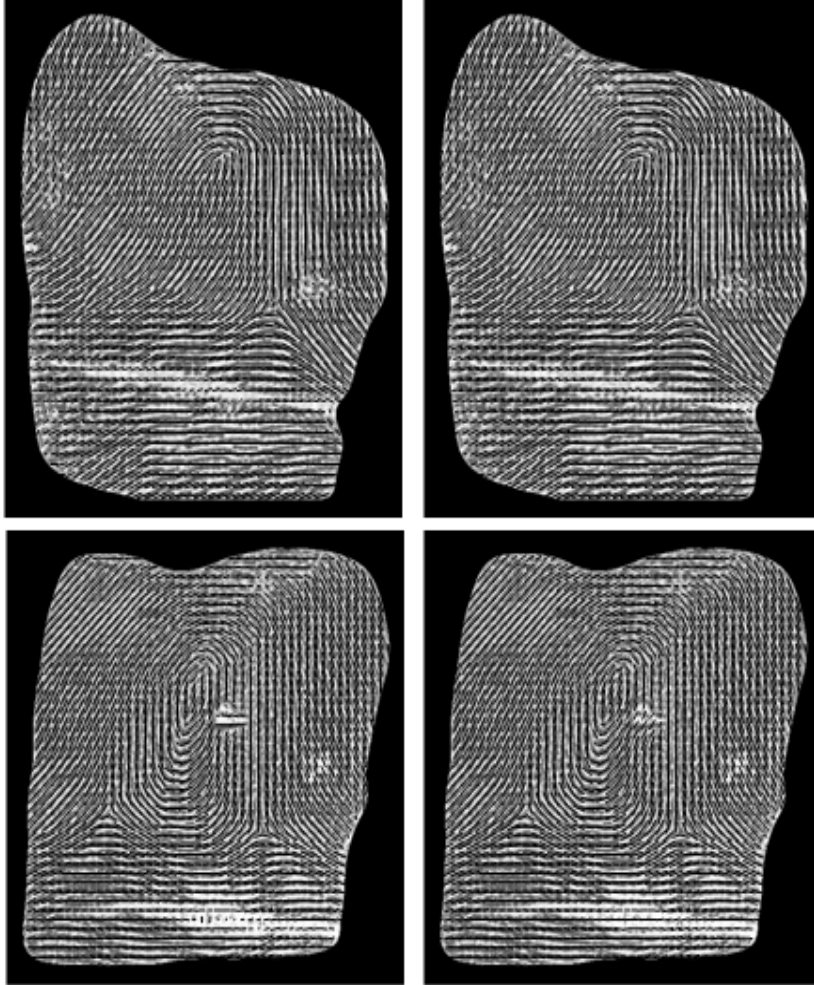


Figure 3. Two fingerprints from NIST Special Database 9 exhibits the orientation regularisation performance. In the first column the orientation extraction outcome is shown, while the second column presents the results of the orientation regularisation procedure.

Due to the shifting effect of the smoother on loops, the last step must be the application of the adjuster over them; in this way the initial position of loops is recovered and the drawbacks of the smoother are neutralised.

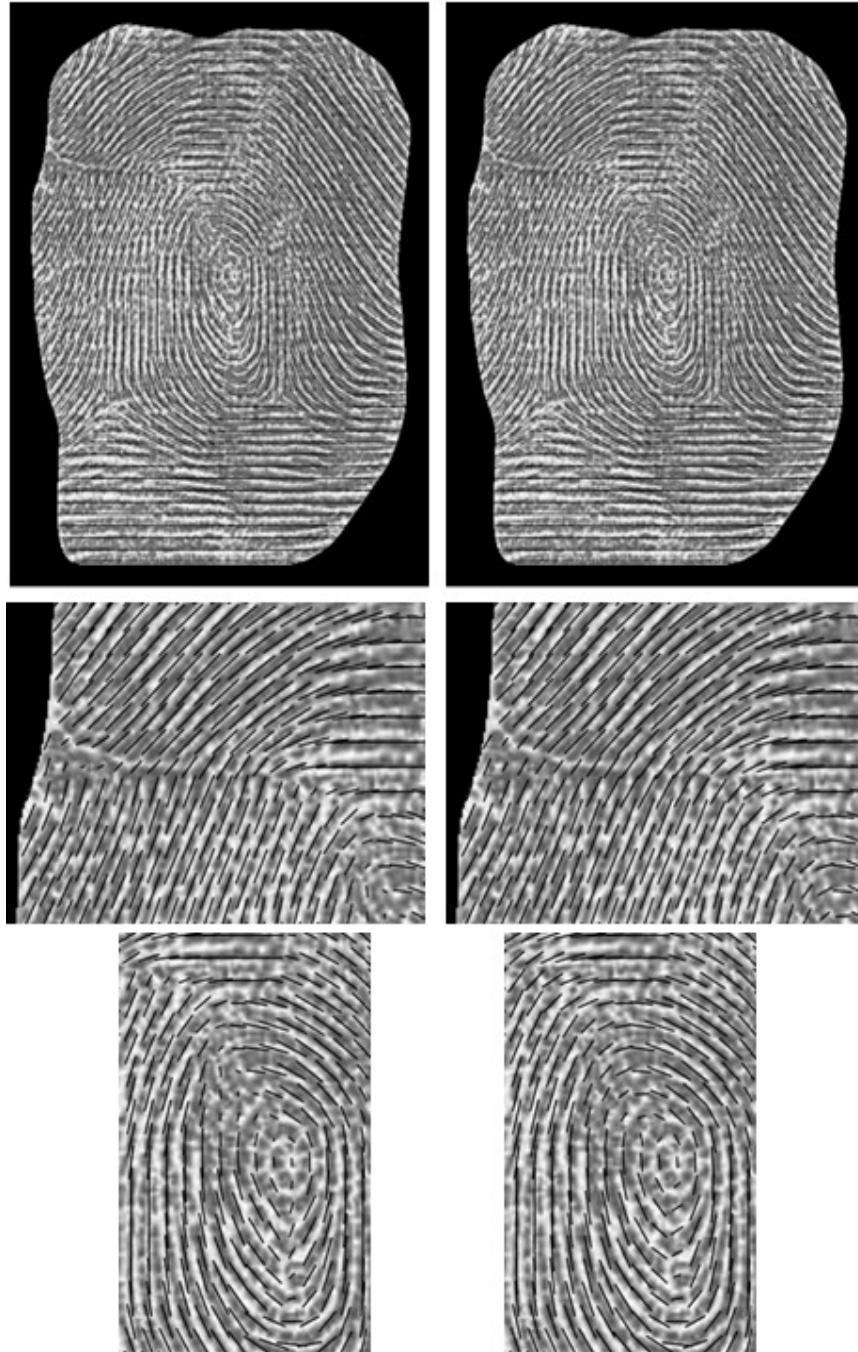


Figure 4. On the left the initial extracted orientation is presented, and on the right the related refined orientation. In the topmost line the whole image is shown, in the second and third lines two magnification of that image are exhibited for better understanding of our algorithm performance.



Figure 5. The effects of regularisation on another fingerprint image: on the left the orientation extraction, on the right the results of the orientation regularisation procedure.

3. EXPERIMENTAL RESULTS

We describe some of the results obtained in the numerical experience with the proposed method. In particular, we first show the outcome of the preliminary steps, such as image equalisation, segmentation, and ridge amplification; then we show the output of the orientation estimation and the orientation regularisation on five fingerprint images.

In Figure 1 we can see the effects of the preprocessing stage, i.e. image equalisation, segmentation, and ridge amplification. The parameters used to generate these images are: $t_L = 0.25$, $t_R = 0.5$, $\tau_V = 2.55$, $t_{M_1} = 0$, $t_{M_2} = 64$, $t_{m_1} = 192$, $t_{m_2} = 230$. The first image of Figure 1, starting from top left, is taken as an input to the algorithm described in Section 2; the second image is the result of the equalisation procedure, the third picture shows the segmentation, and in the last one the outcome of the ridge amplification step is exhibited. Notice the enhancement of ridge-valley structure, a more uniform brightness over the whole image, and the sharp detection of the significant part of the fingerprint image.

In Figure 2 we can see the initial orientation estimation in the top-left image, and the refined orientation in the top-right; due to the small image dimensions, in the bottom line of Figure 2, we exhibit also the magnification of a noisy area where our method performs very well. To compute the orientation estimation we used the parameters $r = 15$, $\sigma_1 = 1$, $\alpha_1 = 2$, $\sigma_2 = 0.85$, $\alpha_2 = 2$, $N_A = 36$, $N_S = 31$, $\tilde{t} = 0.25$. Figure 4 shows the refinement procedure on another fingerprint using the same parameters: in the top line of Figure 4 there are the initial extracted orientation field and result of the orientation refinement procedure on the whole image; in the second and

third lines there are the magnifications of two meaningful areas, where the good behaviour of our method is clear.

The proposed algorithm performs very well in areas with a weak signal, such as creases, and in noisy parts of the image: the orientation there reconstructed exhibits high coherence with neighbouring areas. Figure 3 provides a graphical explanation of such a good behaviour on a couple of fingerprint images.

The proposed method presents a good behaviour also around loops: if the loop structure has been weakened during the orientation extraction stage, our refinement procedure is able to recover it; this phenomenon is clarified by the results presented in Figure 5.

4. CONCLUSION

This paper presents a preprocessing procedure, along with a reliable algorithm to extract the orientation field and to improve it.

The proposed method still has to be tested against a database with ground truth information; this will help us tuning the parameters up and knowing their inner relationships; furthermore, testing our algorithm against such databases will let us compute the accuracy and computational time of the proposed method. Many databases can be found over the Internet, for instance the NIST Special Databases, which will be used in future studies to enhance the present algorithm and to extend it taking into account minutiae extraction.

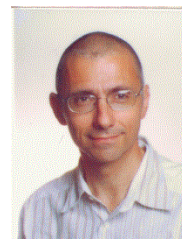
REFERENCES

- [1] Zhang, D. D. (2013). *Automated biometrics: Technologies and systems* (Vol. 7). Springer Science & Business Media.
- [2] Biradar, V. G., & Sarojadevi, H. (2014). Fingerprint Ridge Orientation Extraction: A Review of State of the Art Techniques. *International Journal of Computer Applications*, 91(3).
- [3] Maltoni, D., Maio, D., Jain, A., & Prabhakar, S. (2009). *Handbook of fingerprint recognition*. Springer Science & Business Media.
- [4] Guo, J. M., Liu, Y. F., Chang, J. Y., & Lee, J. D. (2014). Fingerprint classification based on decision tree from singular points and orientation field. *Expert Systems with Applications*, 41(2), 752-764.
- [5] Liu, Q., Peng, K., Liu, W., Xie, Q., Li, Z. Y., Lan, H., & Jin, Y. (2014). Fingerprint singular points extraction based on orientation tensor field and Laurent series. *Journal of Central South University*, 21, 1927-1934.
- [6] Ellingsgaard, J., Sousedik, C., & Busch, C. (2014, March). Detecting fingerprint alterations by orientation field and minutiae orientation analysis. In *Biometrics and Forensics (IWBF), 2014 International Workshop on* (pp. 1-6). IEEE.
- [7] Krish, R. P., Fierrez, J., Ramos, D., Ortega-Garcia, J., & Bigun, J. (2015). Pre-registration of latent fingerprints based on orientation field. *IET Biometrics*, 4(2), 42-52.
- [8] Qi, J., Yang, S., & Wang, Y. (2005). Fingerprint matching combining the global orientation field with minutia. *Pattern Recognition Letters*, 26(15), 2424-2430.
- [9] Tico, M., & Kuosmanen, P. (2003). Fingerprint matching using an orientation-based minutia descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8), 1009-1014.

- [10] Kulkarni, J. V., Patil, B. D., & Holambe, R. S. (2006). Orientation feature for fingerprint matching. *Pattern Recognition*, 39(8), 1551-1554.
- [11] Jiang, X., Liu, M., & Kot, A. C. (2004, August). Reference point detection for fingerprint recognition. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on (Vol. 1, pp. 540-543)*. IEEE.
- [12] Zhu, E., Yin, J., Hu, C., & Zhang, G. (2006). A systematic method for fingerprint ridge orientation estimation and image segmentation. *Pattern Recognition*, 39(8), 1452-1472.
- [13] Oliveira, M. A., & Leite, N. J. (2008). A multiscale directional operator and morphological tools for reconnecting broken ridges in fingerprint images. *Pattern Recognition*, 41(1), 367-377.
- [14] Chen, X., Tian, J., Zhang, Y., & Yang, X. (2006, January). Enhancement of low quality fingerprints based on anisotropic filtering. In *International Conference on Biometrics (pp. 302-308)*. Springer Berlin Heidelberg.
- [15] Lee, K. C., & Prabhakar, S. (2008, September). Probabilistic orientation field estimation for fingerprint enhancement and verification. In *Biometrics Symposium, 2008. BSYM'08 (pp. 41-46)*. IEEE.
- [16] Turrone, F., Maltoni, D., Cappelli, R., & Maio, D. (2011). Improving fingerprint orientation extraction. *IEEE Transactions on Information Forensics and Security*, 6(3), 1002-1013.
- [17] Sherlock, B. G., & Monroe, D. M. (1993). A model for interpreting fingerprint topology. *Pattern recognition*, 26(7), 1047-1055.

AUTHORS

Pierluigi Maponi is associate professor in Numerical Analysis at the School of Science and Technology of Camerino University. Besides fingerprint analysis, his interest for image processing also concerns automatic age estimation, biomedical imaging and diagnostics, satellite imaging. Other research fields are numerical linear algebra, inverse problems and applications, computational fluid dynamics, hazard evaluation for water-related events.



Riccardo Piergallini is full professor in Geometry at the School of Science and Technology of Camerino University. The main field of his scientific activity is low-dimensional topology. In particular, he is interested in the theory of branched coverings, as a tool for representing manifolds and studying various topological and geometric structures on them. Recently, he started to consider computational applications of topology and geometry, specially the ones concerning spatial modeling, image processing, computer graphics and artificial vision.



Filippo Santarelli is a PhD student at the School of Science and Technology of Camerino University. His research interests are fingerprint analysis and voice recognition.



ADAPTED BIN PACKING ALGORITHM FOR VIRTUALS MACHINES PLACEMENT INTO DATACENTERS

Fréjus A. R. Gbaguidi^{1,3}, Selma Boumerdassi^{1,2}, and Eugène C. Ezin³

¹Conservatoire National des Arts et Métiers / CEDRIC, Paris, France

²INRIA Hipercom, Paris, FRANCE

³IMSP /Université d'AbomeyCalavi, Benin

ABSTRACT

The placement of virtual machines is a permanent routine that determines both performance and energy efficiency within Datacenters. Unfortunately, it is a task whose complexity is fully supported by the common sense of the system administrators who must try different scenarios in order to detect the one that best satisfies the constraints imposed by the environment. Bin packing techniques have been used to address similar issues in other areas such as transportation and mass distribution. We try to apply these methods to the problem of placing virtual machines on the physical servers within Datacenters. Our aim is to evaluate the efficiency of this technique at the optimum distribution of the VM while using the minimum number of physical machines and consequently reduce the amount of energy required for their power supply. The results obtained in comparison with the so called brute force method makes it possible to conclude that the Bin packing techniques could help possible to rationalize the use of the physical resources allocated to the operation of the applications in the Datacenters while preserving the SLA imposed by the clients

KEYWORDS

DataCenter, Energy consumption, Bin Packing, Cloud Computing, Datacenters.

1. INTRODUCTION

Virtualization technologies and cloud computing have emerged particularly during the last five years thanks to the widespread needs of ever faster, more sophisticated and always easier technologies. The formerly known enterprise self-computing disappears gradually in favor of outsourcing IT services and on-demand services consumption that both boost business and improve competitiveness in more complex environments. Early, the efforts of industrial and ICT professional shave turned to the improvement of basic technologies of Cloud Computing and first, virtualization. Considered as the forerunner of the modern computing, the possibility of subdividing a physical machine into multiple virtual servers as shown on fig 1 has generated among researchers appetites still not satisfied. Since that hypervisors have been made available, all manufacturers in the computing world migrated their platform in this new perspective, favoring the growth of many datacenters in which the customized services at an optimal price are now possible. The proliferation of data centers around the world, however, reveals the ever increasing needs of material and energy resources for their operation given the demands of the nonstop" generated outsourcing needs and the exponential growth of traffic volumes of online services.

The question, that is the more worrying is that the energy needs of Datacenters now rival the most energy-intensive industries such as aerospace, automotive, to name a few. Many studies have been initiated to try to reduce the magnitude of the problem then generated. First, the hardware manufacturers have invested in the development of components with high electrical efficiency. Then, other software adjustment techniques of the server operating system helped to maximize energy savings. However, in order to respect the constraints of availability and cloud computing services quality, engineers require to size the virtual servers with comfortable margins for avoiding congestions and saturations of processors and memory resources. Energy losses caused by that operation of resources oversized become a new nightmare to which many studies trying to answer. Several algorithms were then developed to optimize virtual machine allocation plan on physical servers. Overall, these algorithms are based on solving a bin packing problem in multiple dimensions with different variants. Two main methods of solving these types of NP-hard problems clash.

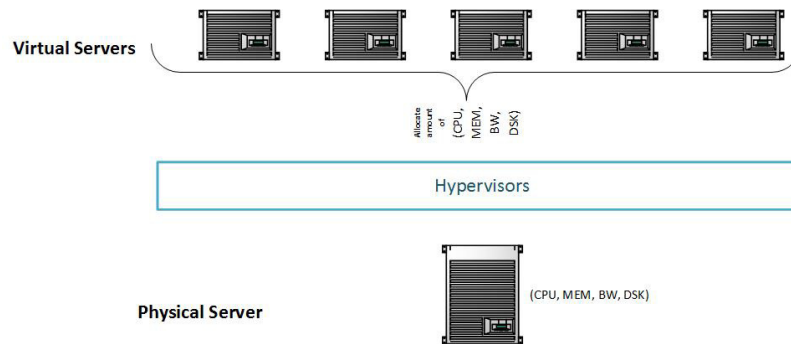


Figure 1. Virtualization classical scheme

On the one hand the so-called "bruteforce" causing very high computation time and incompatible with the process of rapid decision making in the field of cloud computing and on the other hand the heuristic method that combines some logic to find the most realistic solution within an acceptable time. However, the placement problem of virtual machines in the datacenter must take into consideration certain others conditions such as incompatibilities, compliance of combined value memory-processors to a minimum and respect of Service Level Agreements (SLA) constraints. Here we propose a new algorithm that is based on heuristics and produces an occupancy plan of virtuals machines in the datacenter by minimizing the overall processors amount and therefore improve energy efficiency within datacenters. This techniques also help to reduce the decision time for datacenters engineers comparing to manual or brute-force methods.

The contribution of our work is about, on the one hand, the use of the Bin packing algorithm for the design of initial allocation plan of the virtual machines and other hand the introduction of the compatibility constraint regarding to the high availability and data security architectures implemented within Datacenters. The rest of the article is organized as follows: in section 1 we present the state of the art on VM placement techniques, and then describe the Bin Packing tools used in the section 2. We will present our placement approach in Section 3, and the presentation of simulation results in Section 4. Finally we'll conclude the paper.

2. RELATED WORKS

The virtual machine placement algorithms within Datacenters are purchased different goals. Based on several survey [9],[12], [7], [10] [6] the most common one is the reduction of the amount of physical resources deployed and thus optimizing the overall energy consumption. The techniques developed are largely based on heuristics or meta-heuristic of the well-known Bin

Packing problem solving but other techniques such as vector and matrix calculations or methods of artificial intelligence as ant colony systems or Bayesian systems are also used. Many papers like [3] presents a variant of Bin Packing known as the snapsack problem developed to solve de placement problem.[13], [4] proposes an algorithm based on ant colony systems to develop a distribution function of the triplets (cpu, mem,bw) corresponding to the needs of the triplets VM (CPU,MEM, BW) corresponding to available servers. In this article, the utilization rate criteria different physical resources are taken into account in respect of decision making. Similarly, others techniques are proposed by [1], [11] and are based on an iterative approach of VM placement in purpose of cloud service providing. In a more improved version closer to the needs of real data centers, [2] conduct a prior classification of servers resource consumption profile before deciding their distribution on physical servers. This work helps distribute virtual servers to limit sudden later resources congestion on physical servers. The various techniques presented above using various mathematical tools. Techniques based on binpacking problems are most common in the literature and at the same time those that produce the best results. However the classification proposed by [2] does not necessarily correspond to the stresses that are usually found in data centers. We will try to add a manual classification of the servers according to their compatibility.

3. BIN PACKING PROBLEM

Given a set of objects of rectangular shapes of any known size and given a larger rectangular form bin of known dimensions, the bin-packing problem (BP) is to determine the minimum number of bins needed for store without overlap all of these objects (objects do not extend beyond the bins and do not overlap) [8], [5], [6]. More formally, the binpacking problem(BP) is defined as follows: given a set of n rectangular objects $A = \{a_1, \dots, a_n\}$ and an unlimited number of identical rectangles (the bins) of larger dimensions than those of the objects, the problem is to determine the minimum number of bins used to store all objects without overlapping. The problem of bin packing can be approached according to different aspects that we expose below

3.1. Different cases of bin packing problem

Depending on the shape of the treated objects, the binpacking problem can occur in one, two or multi-dimension

- 1BP (one dimension): Is to minimize the number of one dimensional containers (bins) needed to store a list of items characterized by their length
- 2BP (two dimensions): Is a generalization of 1BP. This is to minimize the number of large identical rectangles(bins) needed to store a rectangular shape item list. The items must be arranged in such a way that the sides of the rectangles are parallel to those of the bins

B. Different types of complexity

The problem of storage can be complex depending on the sometimes irregular forms of items or other types of constraints imposed by the problem

- objects of homogeneous or heterogenous forms;
- objects of uniform or different sizes;
- deformable or non-deformable objects

- the number of dimensions of the problem;
- have a single bin (decision problem or maximization problem);
- seek to minimize the number of bins to use;
- seek to minimize the surface or the volume of objects to be placed;
- linear constraints between objects;
- constraints on the order in which objects must be removed from the bin;
- orientation constraints of an object;
- weight constraints (eg the weight of a full bin can not exceed a certain limit)
- Investment constraints, some very heavy objects should be placed at the bottom, other vulnerable must be placed above;
- orientation: objects can be fixed orientation (one speaks of the oriented case) or they can be rotated 90 degrees(the undirected case)
- The guillotine constraint: If it is imposed, it must have the possibility to return the items stored by end cuts parallel end to the dimensions of bins;
- Incompatibility : Some objects can not coexist in the same bin;
- etc.

C. Different solutions

Solving bin packing problems can be very complex and expensive in terms of computing time when it comes to exact solutions. There are fortunately heuristic solutions to approach the problems by providing reliable and very fast results

- Approximate methods: no optimal solution but rapid resolution. Choice of algorithm depending on the problem
- Heuristics algorithms 1 BP
 - Next Fit: assignment of the current object has in the current box if wishes. Otherwise, closing the current box, opening a new one
 - First Fit: allocation of current object has the available box
 - Best Fit: assignment of the current object in the best box
 - Worst Fit
 - Any Fit
- Heuristics algorithms 2 BP

- In one phase directly arrange the objects in the BIN
 - In two phases: preliminary resolution of strip-packing problem (storage of all objects in a bin without height limit) starting by order of Decreasing height before storing in the bin; and apply one of the above methods. We'll talk about NFD, FFD or BFD
- Metaheuristics
 - Several others methods like taboo Search or LowerBounds
 - Exact methods: due to the NP-hard features of Binpacking problem, they resolution time by exact methods is too long and expensive. Facing the lack of a universal algorithm, linear programming is often use to solve problems

4. VM PLACEMENT PROBLEM

4.1 Stating the VM placement problem

Distribute virtual machines on physical servers may seem at first sight a simple division operation of the physical server resource capacity by the same amount of those resources required by virtual machines. Number of data center administrator is exercised manually to resolve the issue by trying to the best of their ingenuity to efficiently allocate resources to VMs. However, with the big size datacenters, it is essential to find tools on the one hand to automate this task and streamline the allocation of resources on other hand. To do this, the distribution of resources is generally reduced to a bin packing problem in which we compute the quantities required for each n resources of the objects to be stored for this dimension while the size of said resource on the physical server is the storage bin. The dimensions N could be represented by trying upto solve the problem in one time for all the n dimensions but we go here to a resolution per dimension. This means finding, based on CPU demands of virtual servers, the number of physical servers required to house them. The standard form of mathematical expression of the problem is as follows.

P is the set of physical servers,

V all virtual machines to distribute,

R_{ij} the amount of resource i on V_j

T_{ij} , the total size of the resource i on P_j

This is to find an allocation matrix M with terms $X_{ij} = 1$ if VM i is placed on the PM j and equal 0 otherwise.

$$\text{Alloc} = \text{Min}(P) \text{ and } X_{ij} = 1$$

This matrix is made with the following conditions:

- 1) The amount of the resource for all VMs placed on a given server can not exceed the total size of the resource on said server
- 2) All VM have to be placed. Thus, the sum of $X_{ij} = 1$ whatever i

3) Allocate the VM on the optimal number of PM. This means finding the minimization function that gives the smallest possible result from a set of solutions.

We add to this classic problem a clause of in compatibility between certain VM. Specifically, it is to solve the technical constraints imposing a distribution of redundant servers in a cluster on different physical machines so that in case of failure of one, continuity of service is automatically provided by the other. To do that we will implement a pre treatment which is to classify the servers depending on their compatibility.

4.2 Our VM placement approach

The classic problems of Bin Packing typically focused on Bin of identical size. We consider, for simplicity, that all physical servers have the same technical characteristics.

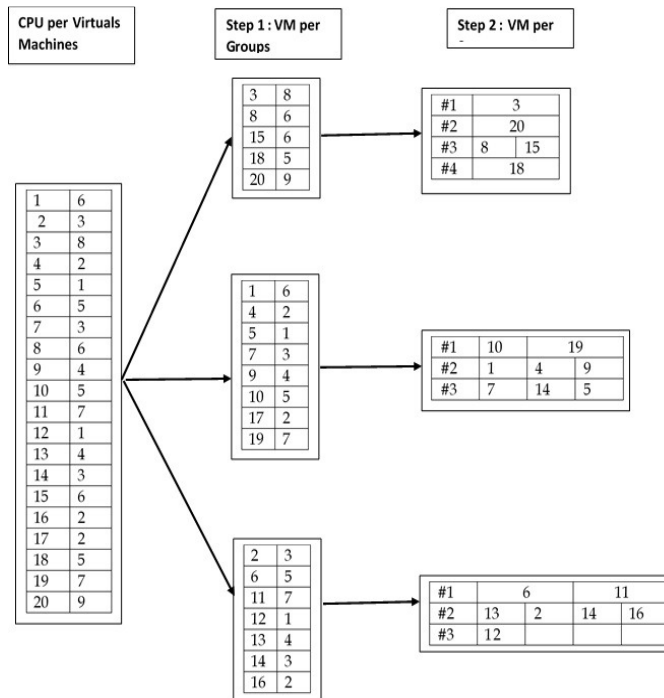


Figure 3. Our Placement steps

We divide the problem into two parts which are:

- Classification by compatibility group virtual servers This is handled semi-automatically. An operator must manually register the virtual machines within a given group. It will be possible to automate this task only through a form to fill out to keep each VM on it compatible group. Then, an algorithm can be implemented for compatibility checks. However, the decision to separate virtual servers can be efficiently taken by an engineer with knowledge of the security constraints, and continuity of service required in Datacenter environments to match the SLA.
- Determining the number of physical servers required per group and consolidation of virtual servers within physical server (see figure 3) Our approach at this level is based on an adaptation of the method of the FirstFit Decreasing (FFD) for allocation of virtual machines on physical servers. Heuristic FFD is resolved into two phases namely: The first phase is to store all resource values required by the VM in a descending order. Then

the First Fit algorithm is applied to successively add VM to physical servers until the limit allowed is reached in a second phase. Thus, VM are stored on the servers to the overall allocation of the list of hosting queries

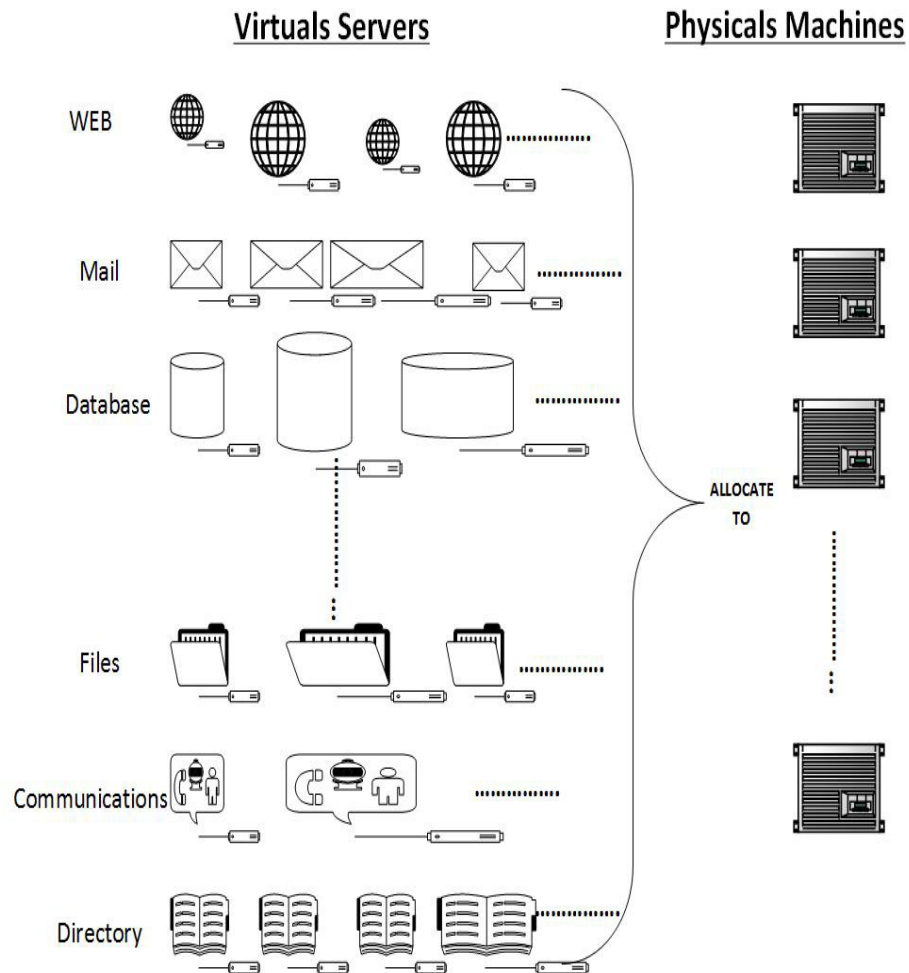


Figure 2. VM Placement Problem

The algorithms of the First Fit Decreasing and Best Fit Decreasing strongly compete in solving similar problems. While in the BFD an optimality criterion is added for the selection of Bins, FFD proceeded to the allocation bin after bin only. Assuming all Bin are identical in size, this additional step implemented in the BFD does not add any efficiency. It's then useful to choose the FFD to save in terms of processing time and thus the operation performance. Since datacenters requires real time operating, the resulting processing time saving fully justified the choice of FFD.

4.3 Placement algorithm

The algorithm is divide into severals procedure including Brute Force, First Fit Decreasing, which compute each one the VM placement. The main program is presented below

```

print("Enter the number of category");
    scan (CAT) ;
    int cat = temp.nextInt();
    for(int j = 0 ; j < cat; j++){
        print("Enter the number of VM");
        scan (n);
        print("Enter the processor length of each VM");
        in = 0 ;
        for(int i = 0 ; i < n; i++){
            scan(val) ;
            in = concatenate(in, val);
        }
    }
print(in);
    print("Enter the processor length of each physical server);
    scan (p);
        bf = BinPackingBruteforce(in, p);
        ffd = FirstFitDecreasing(in, p);
    }
}
    long startTime;
    long estimatedTime;
    startTime = System.currentTimeMillis();
    print("needed bins (" + algoName + "): " + algo.getResult());
    algo.printBestBins();
    estimatedTime = System.currentTimeMillis() - startTime;
    print("in " + estimatedTime + " ms");
    print("\n\n");
}

```

Figure 4. Placement Algorithm

5. EXPERIMENTS AND RESULTS

5.1 Experiments framework

We try to demonstrate through simulations the performance of our approach to solving virtual machine placement problem by comparing an implementation of FFD with a brute-force method. The problem of Bin packing is NP-Hard, then the resolution by the brute force method consists on going to each occurrence of virtual machine and then determine the ability to allocate or not to the physical machines.

Our study framework is one of the institutional Datacenters in Benin in which the several public administrations regularly requests hosting their application. It hosts at ually a hundred of servers but we will limited our study of a dozen HP Blade BL 350 with identical specifications in purpose to meet the constraints of the Bin Packing problems. Servers characteristics are as follows:

- Processor: Dual processor with six cores at 2,4 GHz eachper server
- RAM memory :128 Gbytes
- Hard disk : this resources are use over the SAN systemwith over 100 TBytes of capacity
- Network Controller: 10 Gbps

To facilitate data processing into our program, we voluntarily libellons the amounts of such resources unless decimals parts. Virtual machines on it is requested through a hosting application in which customers can complete their needs among other, amounts of CPU, Memory capacity, HDD capacity and number and speed of network controllers. Here again, we limit ourselves to twenty virtual machines requests. The summary of the resource requests of our experiments can be as follows:

Table 1. SERVERS LIST BY GROUP.

| N° | CPU Request | Group |
|----|-------------|-------|
| 1 | 6 | 1 |
| 2 | 3 | 1 |
| 3 | 8 | 1 |
| 4 | 2 | 1 |
| 5 | 1 | 1 |
| 6 | 5 | 1 |
| 7 | 3 | 1 |
| 8 | 6 | 2 |
| 9 | 4 | 2 |
| 10 | 5 | 2 |
| 11 | 7 | 2 |
| 12 | 1 | 2 |
| 13 | 4 | 2 |
| 14 | 3 | 3 |
| 15 | 6 | 3 |
| 16 | 2 | 3 |
| 17 | 2 | 3 |
| 18 | 5 | 3 |
| 19 | 7 | 3 |
| 20 | 9 | 3 |

The machines are virtualized using VMware hypervisorESXi 5.5. The aim of our simulation is to compare the number of currently used server in the datacenter using a manual method and Brute force with that obtained using our adapted version of First Fit Decreasing algorithm. The simulation is to enter into the program developed for this purpose, the needs of users of Datacenter and the available quantity of the resource at the server. The simulation output is to determine for each method, the number of necessary physical server as well as the processing time.

5.2 Results

In the table II, we presents for each VM packing method the total duration of the process and thee number of physical Machine (PM) required to host the groups of VM. For manual technique, the duration is marked as Non Available since it depend on the ability of each Datacenters engineers.

Table 2. ALGORITHMS PERFORMANCE

| Methods | Group 1 | | Group 2 | | Group 3 | |
|-------------|---------|----------|---------|----------|---------|----------|
| | PM | Duration | PM | Duration | PM | Duration |
| Manual | 4 | NA | 4 | NA | 4 | NA |
| Brute Force | 4 | 0 | 3 | 875 | 3 | 31 |
| FFD | 4 | 0 | 3 | 0 | 3 | 0 |

Considering the results, we clearly see that the determination of the initial VM placement scheme is optimal by the method of Bin Packing. The evaluation also shows that the calculation time by the method of Brute-force is very high compared to the FFD at a very high factor. The preliminary classification of machines by compatibility group is also an important feature of our study as this reality is often ignored in many similar work even though it is a basic rule of VM placement within reals Datacenters Remembering that our sample is small for simplicity, it is useful to note that the allocation scheme determination duration can be several times higher when considering a large size Datacenters (hundreds or thousands of servers). The use of FFD based technique is essential to mitigate that weakness and respond effectively to the flow of hosting applications requests in the datacenter. Determining an optimal placement scheme of virtual machines in a data center represents only the beginning of the placement spots since the quantities of resources requested initially are very different to that consumed after the production start leading huge loss of resources. These under consumption result in a waste of electrical power because once setting up, VM consume nearly the amount of energy that they need when they are fully loaded. To overcome this, a periodic reallocation of virtual machines on physical servers is necessary. However, even if the methods of Bin Packing can always be helpful, other algorithms based on appropriate criteria are further necessary.

6. CONCLUSION

The optimization of energy consumption within the datacenter is an area where research is so far very fruitful. The key lies in virtualization, technical development to improve the electrical efficiency of deployed servers and better allocation of virtual servers. With constantly growing needs of cloud computing, the placement problem of virtual machines on physical servers deployed in data centers has become central. It is divided into two parts which are: i) the initial placement scheme and ii) the optimal reallocation based on the real needs of users over time. Our goal in this work was to propose an approach to the issue of initial placement through methods of Bin Packing. Thanks to results produced in other areas like logistics, algorithms based on bin packing carry promises for improving resource consumption on servers within datacenters. In fact, our tests on a sample taken from a real environment show compared to other traditional methods of VM placement on servers that the heuristic of the First Fit Decreasing reduces so sensitive waste of resources on physical servers. Added to this, the optimization of the time required for determining the proper placement scheme is also highlighted. This reduced the time for Datacenters administrators in intelligence decision namely for customer VM hosting since cloud computing technologies requires more responsiveness and more efficiency for better satisfaction. The results, however, are just an initial basis for the resolution of energy efficiency issues in the datacenters. The increasing size of these, the variety of constraints related to technology and customer requirements involve the development and validation in varied environments of other techniques in the objectives of dynamic management of VM allocation

under production. Existing virtual machine migration methods must be improved by combining existing solutions with the Bin Packing tools.

REFERENCES

- [1] MB Arya and Ajay Basil Varghese. A combined bin packing vm allocation and minimum loaded vm migration approach for load balancing in iaas cloud datacenters.
- [2] Norman Bobroff, Andrzej Kochut, and Kirk Beaty. Dynamic placement of virtual machines for managing sla violations. In *Integrated Network Management, 2007. IM'07. 10th IFIP/IEEE International Symposium on*, pages 119–128. IEEE, 2007.
- [3] Ricardo SteghCamati, AlcidesCalsavara, and Luiz Lima Jr. Solving the virtual machine placement problem as a multiple multidimensional knapsack problem. *ICN 2014*, page 264, 2014.
- [4] Eugen Feller, Louis Rilling, and Christine Morin. Energy-aware ant colony based workload placement in clouds. In *Proceedings of the 2011 IEEE/ACM 12th International Conference on Grid Computing*, pages 26–33. IEEE Computer Society, 2011.
- [5] Richard E Korf. A new algorithm for optimal bin packing. In *AAAI/IAAI*, pages 731–736, 2002.
- [6] Andrea Lodi, Silvano Martello, and Michele Monaci. Two-dimensional packing problems: A survey. *European journal of operational research*, 141(2):241–252, 2002.
- [7] Fabio Lopez-Pires and Benjamin Baran. Virtual machine placement literature review. arXiv preprint arXiv:1506.01509, 2015.
- [8] EG Co man Jr, MR Garey, and DS Johnson. Approximation algorithms for bin packing: A survey. *Approximation Algorithms for NP-Hard Problems*, pages 46–93, 1996.
- [9] ZoltánÁdám Mann. Allocation of virtual machines in cloud datacenters—a survey of problem models and optimization algorithms. *ACM Computing Surveys (CSUR)*, 48(1):11, 2015.
- [10] Kevin Mills, James Filliben, and Christopher Dabrowski. Comparing vm-placement algorithms for on-demand clouds. In *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*, pages 91–98. IEEE, 2011.
- [11] Weijia Song, Zhen Xiao, Qi Chen, and Haipeng Luo. Adaptive resource provisioning for the cloud using online bin packing. *Computers, IEEE Transactions on*, 63(11):2647–2660, 2014.
- [12] B Benita Jacinth Suseela. Survey on vm placement algorithms. *International Journal of Engineering Trends and Technology (IJETT)*, 6(7):349–352, 2013.
- [13] Gaochao Xu, Yan Ding, Jia Zhao, Liang Hu, and Xiaodong Fu. A novel artificial bee colony approach of live virtual machine migration policy using bayes theorem. *The Scientific World Journal*, 2013, 2013.

AUTHORS

Fréjus A. R. GBAGUIDI is a PhD student in Computer science at Cnam Paris and Abomey-Calavi University Benin (IMSP-UAC). With a Master Degree in Network and Information System, He's a Datacenter and network specialist with many skills in several fields including Information System Security and Audit.



Selma BOUMERDASSI is Associate Professor, at Conservatoire National des Arts et Métiers (CNAM) Paris in security and localization in ad-hoc networks and also Researcher at Institut National de Recherche en Informatique et Applications (INRIA-France)

Eugène C. EZIN is a Lecturer of the universities of CAMES, Director of Institut de Formation et de Recherche en Informatique at Abomey Calavi University. His research fields include Artificial Intelligence, Neural Networks, Multimedia and Computer Software



CRYPTOGRAPHIC STRENGTH ESTIMATION USING SPURIOUS KEYS WITH CONSIDERATION TO INFORMATION CONTENT IN THE MESSAGE

Mekala Rama Rao¹, L Pratap Reddy², BHVS Narayana Murthy³ and
Maruti Sairam Annaluru⁴

^{1,2}Department of Electronics and Communication Engineering, Jawaharlal
Nehru Technological University Hyderabad, Telangana, India

^{3,4}Research Centre Imarat, Hyderabad, Telangana, India

ABSTRACT

Among the available private key cryptosystems, namely stream ciphers and block ciphers, the advantage of block ciphers is that they can be synchronized i.e. losing one ciphertext can not affect the correctness of the decryption of the following blocks. The encrypter used in block ciphers is a memoryless device. Block ciphers can be easily standardized due to the fact that they transmit information in blocks. But the disadvantage is that identical plaintexts result in identical ciphertexts. These data patterns are not hidden by the algorithm, resulting in higher influence of cryptanalysis process. Strength of block ciphers is exposed mainly into the exploration of weakness of cryptosystem. Barring this approach, strength estimation based on spurious key analysis is proposed in this paper.

Till recent period, strength of a Cryptosystem is identified with the increasing key length. However, as per Shannon's proposal, strength of a Cryptosystem is dependent on Message also. Depending on the length of the message and the message space, we can estimate the actual strength of a Cryptosystem. As part of Shannon's model, spurious keys is the concept adopted for identifying the strength of the Cryptosystem. Standard block ciphers; ARC2, Blowfish, CAST, DES; are evaluated to understand Shannon's principle of Information Theoretic approach using Spurious keys. Spurious key generation algorithm is designed, developed for evaluating the strength of Cryptosystem. Spurious key logic and Key scheduling logic are the two main blocks of the proposed approach. Behavior of Spurious keys is evaluated on message text of two languages, through selection of ten different sub key spaces. Each sub key space is independent of other and is constructed with 10^8 keys from the total key space of 2^{64} . It is observed that the number of spurious keys identified in each sub key space is almost close to similar value of the respective language. Comparison is made through this evaluation to explore algorithmic strength with that of computational burden of the algorithm, which will help selection of algorithm based on the critical requirements of the field. The very purpose is to examine the possibilities of considering spurious key analysis as one of the strongest methods to estimate the strength of a Cryptosystem. Spurious key analysis is performed on two sets of plaintexts containing two different scripts namely English and Devanagari.

KEYWORDS

Cryptography, Spurious Keys, Complexity, Strength of Algorithms, cryptanalysis, Language Based Security

1. INTRODUCTION

With the increased digital access in the areas of communication and financial transactions, shared data is becoming more and more personal in nature. The increase in personification of data is more significant these days. Huge amount of personal data is being shared every second. In this era of digital transactions, the security of one's wealth depends on the cryptosystem used by the e-wallet provider. The privacy of a social networking user depends on the strength of the cryptosystem used. The ever increasing growth of communication networks and the emergence of Internet Of Things (IoT) demands for undesired disclosure of information.

Present era of information explosion through social networking, it is undesirable to rely on concealment systems [3]. Information-theoretic analysis of information hiding [4] suggests that the capacity of data hiding suffers issues like side information being available to the intruder, tapping of wires, low rates of reliable transmission etc. Ideally, the system being used to transmit information should provide security even when the process of transmission is kept open. Cryptosystems are required to serve this purpose. The electronic data need to be maintained confidential for a long period. But security systems cannot provide long term security due to the increasing computational power. In the present trend of cloud computing, the owner of the data may not be confident that no adversary has access to the data, as it is being transmitted through a public network. If the original data is accessed, copied and stored then re-encryption of the data cannot guarantee any further security as the data is already stored. Deletion of data may not be completely guaranteed. Moreover, if quantum computers are used, then computing factors of large numbers and logarithms are also feasible, where RSA algorithm may not guarantee security.

Knowing the strength of a cryptosystem is vital in many applications where secure communication is desired. Cloud computing emphasized the need for data security [9]. Intruders can identify the vulnerabilities of a user whose security can be easily exploited, in such cases security of the entire cloud or network is at risk. Poor user security behaviour is a significant, perhaps even the major, detriment of the level of security incidents a company suffers [10]. Using a secure cryptosystem and efficient key management processes are necessary for such environment as most of the e-commerce applications are being run by cloud [11]. Key Dependent Messages (KDM) is another concept of encryption which allows requested plaintexts to depend on the underlying decryption key [36]. This concept is tested as an attack on block ciphers. This attack is called key dependent input (KDI) attack and it was found that for every function, a KDI secure encryption scheme can be built [37]. Encryption scheme of key dependent message (KDM) secure is reported to be secure even against an adversary who has access to encryptions of messages that depend on the secret key [38]. With ever increasing attacks on personal data and banking system through cyber space, it is important to have new dimensional measures to estimate the strength of a cryptosystem. One such method which considers message as a text and measures the strength of the system based on the language of the message, is presented in this paper.

1.1. Secrecy System Overview

A secrecy system, in general, can be viewed as a transformation from message space to ciphertext space [1]. This transformation should be reversible in nature to assure unique mapping between plaintext and ciphertext. Though encryption and decryption is same, the unique mapping between

these two spaces is determined by the chosen key. In these private key cryptosystems, sender and receiver should share a secret key which is unknown to the intruder. However, construction of algorithm, chosen key and the construction of message provides apriori knowledge to the intruder. The process of cryptanalysis involves obtaining posterior probabilities of a cryptogram from the set of keys and messages that are assumed to be associated with that cryptogram. The cryptosystem which provides less posterior probabilities from the apriori knowledge is treated as a strong cryptosystem.

The security of a cryptosystem is considered to be based on complexity theory [2]. The complexity of a cryptosystem is either algorithmic complexity or key complexity. When a cryptosystem is complex, it takes more time for the cryptanalyst to break it. The system can then be considered as secure till that time i.e. computationally infeasible. Security of these systems lies in the fact that brute force attack becomes impractical with the increased size of the key and when the algorithm is complex. It should be noted here that there are methods available which take less time than the exhaustive key search [5]. Linear and differential cryptanalysis on various cryptosystems proved that the success rate of breaking ciphers is increasing with the improved computational speed [6].

1.2. Cryptographic Strength Estimation Review

Cryptosystems are considered to be secure by the fact that the construction of S-boxes is non-linear in nature [7]. Under the additional hypothesis that these S-boxes constitute overdefined system of algebraic equations, XSL attack was performed on Serpent and it was proved that the security of ciphers does not grow exponentially with the number of rounds [8]. Way back in 1977, it was suggested that an exhaustive key search on parallel machines can break the NBS data encryption standard. The algorithmic complexity of a cryptosystem is achieved by iterating a weaker function in various rounds. It is later proved that DES can be reduced to 8 rounds and can be broken in less than 2 minutes and the complexity does not grow exponentially [12]. If DES is reduced to 15 rounds, then it is breakable faster than exhaustive search. Differential cryptanalysis was successful on FEAL cryptographic algorithm also [13]. The attack is based on chosen plaintext attack where cryptanalyst has a bunch of plaintexts to apply on a known algorithm and obtain corresponding ciphertexts. It is a black box concept where the cryptanalyst chooses two particular plaintexts whose differential value is known and calculates differential value of the corresponding ciphertexts. After analyzing several such plaintext–ciphertext combinations, a relation is derived between plaintext – ciphertext pair, there by exploiting the complexity of the algorithm. The security of optical cryptosystems is also exploited using known plaintext attack [14]. IDEA cryptosystem was also broken using an advanced differential attack, namely narrow biclique method which uses meet-in-the middle attack [15]. Though the complexity of this cryptanalysis is more, it is proved that even the complex algorithms can be broken using advanced computational systems. In 2012, cryptanalysis was demonstrated on full AES also [16]. AES is successfully cryptanalysed within 8 rounds. Though cryptanalysis of higher rounds was not successful for AES, 9th and 10th round cryptanalysis was successfully performed on AES like systems [17].

Many researchers formulated various cryptanalysis methods over the years and are successful in breaking ciphers up to few rounds. UCL crypto group has summarized almost all the cryptanalysis methods available during 1990s [41]. If the security of a cryptosystem lies in the nonlinearity of its algorithm, S box more specifically, then identifying the near linearity is the key factor of linear cryptanalysis. In 1994, linear cryptanalysis method was proposed for DES [18]. It is performed on SPECK [19], reduced round SIMON [20], FEAL [21], Serpent [22], PRESENT [23] and AES [24] algorithms in recent years. Biclique cryptanalysis is also demonstrated on MIBS-80 and PRESENT-80 block ciphers [25] and AES [26] algorithm. It is

reported that the vulnerability of DES to bruteforce attack can be overcome by the 64 bit key cryptosystem, Blowfish [27]. Several investigations are performed and suitable key lengths are proposed for various cryptosystems [28]. Later, the use of biometric [29], face hashing [31], face recognition [32], hand written signature [33] and voice [30] based key generation was suggested by several authors. But with the known plaintext attack, the information about the key can be obtained. The cryptanalyst obtains several plaintext-ciphertext pairs which are generated using the same key. On the basis of this a priori knowledge, the key is determined for use in reading later cryptograms for which he need not know the plaintext. The key factor in complexity of increased key length is that the time taken to apply all possible keys becomes infeasible when the key length is increased. But, the interesting fact is that the cryptanalyst actually needs the sub keys, but not the actual key to break the cipher. Frequent changing of the key is considered to increase the security. The security may be increased by frequent changing of the password but the advantage is small [39]. This is because, even in exhaustive search, the attacker is expected to have a successful guess halfway only through the search i.e. if the total number of keys is K , then only $K/2$ keys are sufficient to guess the actual key. Key recovery was practically performed on 10 rounds AES algorithm [40], which is considered to be the strongest cryptosystem.

Linear cryptanalysis is a known plaintext attack that uses linear relation between inputs and outputs of an encryption algorithm that holds with certain probability. This approximation can be used to assign probabilities to the possible keys and locate the most probable one. Once linear approximations of the S-boxes are found, the problem is to find a way to combine those individual approximations to establish a final approximation of the cipher that involves plaintext bits, ciphertext bits and key bits only.

Differential cryptanalysis is another powerful method which analyses the differences of ciphertext pairs resulted from plaintext pairs of particular difference. Linear transformations like bit permutations, key additions etc cause differences between two texts. For an S box with n input bits and m output bits, 2^n input combinations are possible and for every input difference, getting 2^m output differences is an a priori knowledge to the crypt analyst. Cryptanalyst examines all these possible differences and tabulates the number of occurrences of each difference. These probabilities are used in identifying the original key. Differential cryptanalysis also requires $r-1$ rounds to be analyzed. Let Δ_p is the plaintext difference and Δ_c is the resultant ciphertext difference. If any plaintext pair with Δ_p probability provides a ciphertext pair with Δ_c probability after $r-1$ rounds, then that pair is called right pair. These right pairs are used in analyzing the behavior of algorithm for multiple plaintexts in order to break it. As a counter measure to differential cryptanalysis, the differential propagation inside an S-box is maintained as low as possible and the number of S-boxes is maintained as high as possible.

Differential-linear cryptanalysis is a chosen plaintext attack where the linear cryptanalysis is used to provide the differential characteristics of the cipher. An 8 round attack against DES recovers 10 bits of key with 512 chosen plaintexts. However, expanding the attack to higher number of rounds is not found in literature. The strength of a cryptosystem is also estimated using side channel attacks like power consumption [34] and throughput [35] of the system. Throughput of a system lies in the algorithmic complexity which can be exploited using linear and differential cryptanalysis. With the decreased power consumption rates in lightweight cryptographic algorithms, the strength comparison requires a new measure than power.

All these attacks, and many more attacks introduced thereafter, are applied on various ciphers either to analyze their behaviour or for key recovery. Increasing number of rounds, increasing size of the key, building more complex S-box structures, using more S-boxes etc. are considered to be counter measures for all these attacks. These counter measures are based on increasing time to analyze a cryptosystem. All these counter measures are actually adding more complexity to the algorithm and consuming more power and space. This increased complexity results in reduced

speed and throughput of a block cipher. The message is also considered as a mere bunch of 0's and 1's. The information contained in the message is ignored. The security that can be provided by the message itself is another important aspect that is yet to be taken care of.

2. CONCEPT OF SPURIOUS KEYS

Many researchers believed that it is important to create uncertainty in original data in order to make it secure [42]. Purpose of S-boxes is for meeting this requirement of creating uncertainty. It is achieved by a principle that if an S-box is complete then its inverse is not complete. It is therefore believed that the key cannot be discovered by applying the inverse of an S-box.

Every message in general posses some amount of apriori knowledge. This apriori knowledge is useful to cryptanalyst. When an encrypted message is decoded using various keys, most of the keys may decrypt the cryptogram in such a way that it can be easily eliminated as a wrong key. The easy elimination of such keys is possible due to the fact that they contain gibberish (non-text characters). The cryptanalyst knows that the message cannot contain non-text characters. Some of the keys will result in text like message after decryption. However, the resultant text like message may or may not posses any meaning. All these keys are the probable keys. One of them is actual key. These keys introduce ambiguity to cryptanalyst and are called Spurious Keys.

Spurious key analysis is not applied in historical measures for cryptographic strength estimation. But the role of redundancy and unicity distance in the non-linearity of a language was studied as a separate entity. The study presented in this paper combines these language aspects with security measure which results in spurious key analysis. Present study concentrates on establishing spurious keys as a vital component for cryptographic strength estimation. Efforts are made to investigate and validate the suitability of this measure. However, this is just an alternative to existing methods when plaintext and ciphertext are considered as mere bit streams. But these texts contain crucial information of the data being sent or received. If that information content is taken into consideration, then this analysis is vital for strength estimation.

2.1. Spurious Key Model

Assuming that the original message needs to be decrypted is a text, by applying a ciphertext only attack on the cryptosystem; all the decryptions do not provide a text-like-text as the output. This provides a scope for the cryptanalyst to eliminate all those decryptions which do not look like text. The search for original decryption can now be narrowed. But, in the process of identifying the original message, the cryptanalyst meets certain set of keys which provide meaningful output. These keys are spurious keys and are not possible to eliminate without knowing some extra information about the message. The primary goal of generating spurious keys is to identify the strength of a specific algorithm being used. The more the spurious keys, the more difficult it is to identify the actual key. All these keys are equiprobable keys where any one of them can be the actual key. The cryptanalyst applies the keys until the satisfactory results are obtained. As all these keys provide a text-like-text as the output, fetching only one key among them is a challenging task. In order to understand the cryptanalyst's perspective, a simulation model can be designed for generation of spurious keys.

The process of generating spurious keys involves two major steps, one is encrypting a message using a specific key and the other is decrypting the ciphertext using random keys for identifying spurious keys based on the decryption. A spurious key generation algorithm is proposed with an assumption that the plaintext is a meaningful text. The flow chart for spurious key generation algorithm is given in Figure 1.

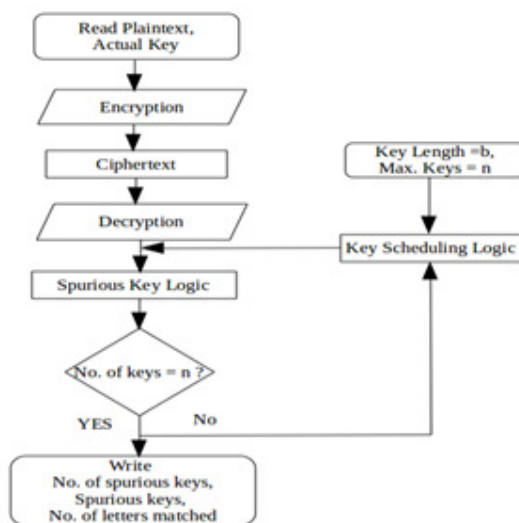


Figure 1. Spurious key generation algorithm

Initially, a ciphertext is obtained by applying the plaintext and a key to the encryption algorithm. The plaintext can be retrieved back by applying the ciphertext and the key to decryption algorithm. A single key is required to obtain the ciphertext from the plaintext, using any cryptographic algorithm. This ciphertext is assumed to be available to the cryptanalyst. The cryptanalyst tries to apply various keys to identify the original message. To do so, the cryptanalyst makes some assumptions like the text might belong to a specific language, the content might be related to a specific task etc.

The encryption and decryption blocks can be of any cryptosystem, like AES, DES, Blowfish, ARC4 etc. These models can be implemented in any programming language like C, Python etc. In this method, encryption is carried out only once, using the actual key. But decryption is performed continuously using various keys. Each key will provide decrypted text which may or may not be a meaningful text. The process of applying various keys is considered as a specific block called key scheduling logic. Another block is called spurious key logic. It is used for identifying whether a decrypted message is text-like-text or not. As the cryptanalyst does not know the actual key being used, he may try to apply various expected keys depending on a priori probabilities. This process can be continued for a predefined time period or for all the possible keys in the range of key size of the algorithm being used. The key scheduling logic is an important block in the process of generating spurious keys. The main objective of this block is to generate keys continuously with the key size specific to the encryption algorithm used.

Due to the limitations in computational capabilities, it is difficult to apply all possible keys to the algorithm and record the spurious keys. Application of all possible keys depends on the key size of the algorithm and the computational capability of the system being used for simulation. For example, applying all possible keys present in the entire key space to AES algorithm is not feasible on a single computer even with the present day high speed computers. Hence, a specific method is required for scheduling the keys. This is to be done with an upper limit on the number of keys. Key scheduling can be done until a specific time period (or) for a specific number of keys. In the first method, the keys are generated and applied to the decryption algorithm for a specific time period and the number of spurious keys is calculated. In the second method, the keys are generated for a specific key space without considering the time.

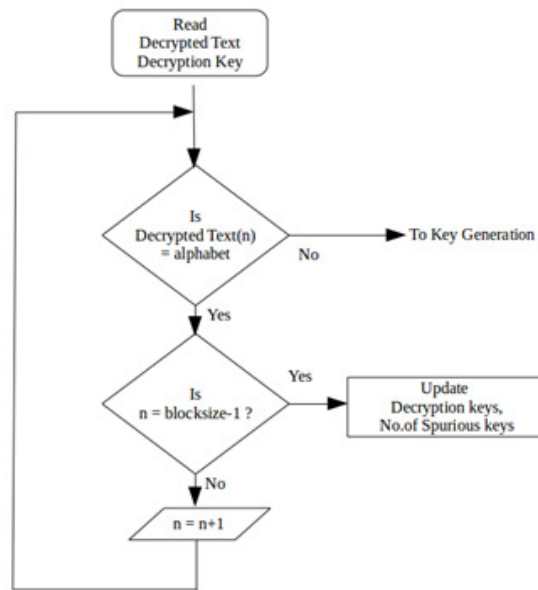


Figure 2. Flow chart for Spurious Key Logic

The logic involved in generating the spurious keys is shown in Figure 2. Spurious keys are those keys which generate a text like text. For example, upon applying a specific key, if the decrypted result is “h d I”, then the cryptanalyst can easily identify that the applied key is not the correct one. The cryptanalyst can eliminate such keys and narrow the process. The key space is defined by cryptanalyst depending on a priori probabilities i.e. the knowledge about the algorithm, language, process etc which can be predicted without actually knowing the actual plaintext, key etc,. The entire process can be continued for a predefined time period or for all the possible keys in the range of key size of the algorithm being used. The main objective of key scheduling block is to generate keys continuously with the key size specific to the encryption algorithm used. All possible numeric keys, i.e. 00000000 to 99999999 are generated using key scheduling logic block which is a simple 8 byte counter in this case.

As the number of spurious keys increases, the process of cryptanalysis becomes difficult. In turn, spurious keys increase security of the algorithm. The total message space contains numeric digits, alphabets and some special characters. Each message data is uniquely mapped to a separate ciphertext data. Due to this one-to-one mapping, a specific ciphertext transforms into the original message when decrypted using the actual key. The interesting fact is that the entire ciphertext space is not limited to the size of message text space i.e. there are many ciphertexts beyond the entire volume of message space. Due to this specificity in any cryptogram, the suspected keys can be reduced to a limited volume i.e. all those keys which provide gibberish can be easily eliminated from the expected key space. This is where the cryptanalyst has an advantage of limiting his search in order to break the code. When compared to exhaustive key search, the computational effort is reduced if the number of spurious keys is less. The strength of a cryptogram can hence be estimated based on its ability to produce the number of spurious keys. If all the keys are spurious then the system is a perfect secrecy system.

2.2 Limitations of Study

Though the paper is intended to establish spurious key analysis as a valuable parameter in estimating the strength of cryptosystems, the study is performed with certain limitations. The major limitation is in terms of key space. Though 264 keys are possible, entire key space is not

used for computational convenience. The algorithms used in this study accepts any ASCII character as key, but only 10^8 disjoint numeric keys are considered. Message size is also limited because of the nature of block cipher where they convert a message of any size into blocks of fixed size. Though, there are several cryptosystems available, only 64 bit block ciphers are considered for evaluation. Another important limitation is that the study is carried out with a presumption that the information content is also important in a cryptosystem.

3. EXPERIMENTAL RESULTS

All evaluations are performed on Intel® Core™ i7-4770 CPU @ 3.40GHz × 8 processor. The python cryptography toolkit is used for validations. Evaluation is carried out with the help of a common key, while varying the message text. Test samples of 50 varied text messages of 4 characters, 6 characters and 8 characters are considered for this evaluation. These 50 samples are taken at random from the dictionary. They hold no specific relation among themselves. The effect of message text size on each block cipher is also taken into account. Encryption is carried out with the key word '12345678'. Even though numeric key is considered, it was found that any key in the key space has the same effect. Decryption is performed on the ciphertext generated from the encryption process using 10^8 key combinations generated from key space of numeric keys of length 64 bits. While performing brute force, the decrypted message text is compared with the text space consisting of English alphabets. If the decrypted message text consists of any symbol other than English alphabets, then the corresponding key is considered as non spurious key. All spurious keys in the range of key space as defined, are listed out as spurious keys. The number of spurious keys observed for each message text is presented in the following section.

A sample spurious key for four algorithms, viz. DES, ARC2, Blowfish, CAST, with the respective text message as well as the decrypted text like message are listed in the table below. The plaintext messages are also listed in the Table 1.

Table 1. Sample spurious keys for English text

| Algorithm | Plaintext | Ciphertext | Text like Message | Spurious Key |
|-----------|-----------|--------------|-------------------|--------------|
| DES | alkaline | □d*□□W□□ | xcKOWAVI | 002c4e02 |
| ARC2 | duck | e2QwfoIOD2M= | QuDIIXoO | 01189996 |
| Blowfish | computer | W1x0/2DbLMQ= | mwethwzo | 00177391 |
| CAST | abacus | LQYdQMLqVU8= | MqgfDgHh | 00124874 |

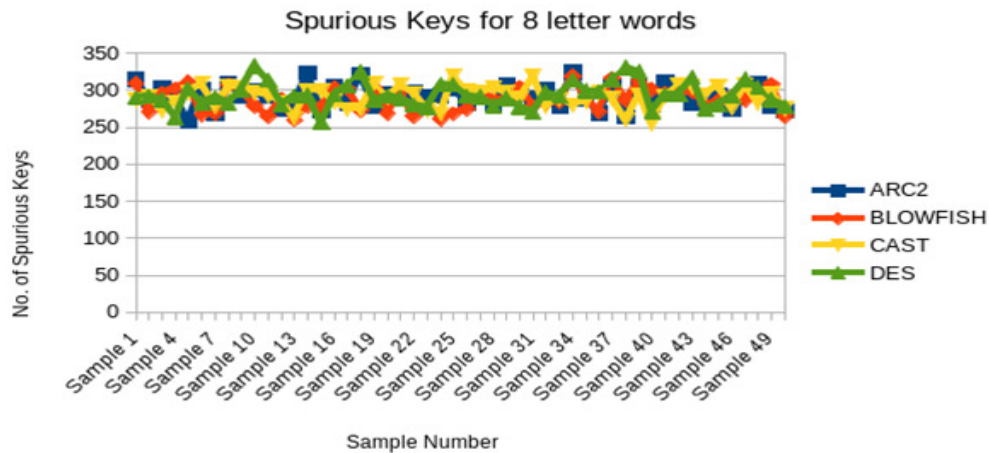


Figure 3. Spurious key analysis for 8 characters English Message

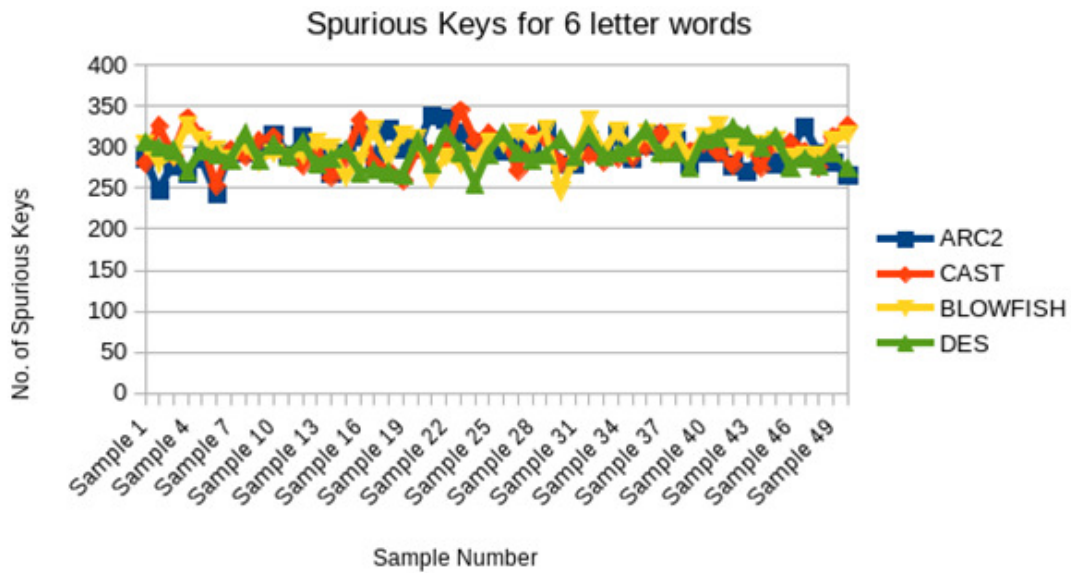


Figure 4. Spurious key analysis for 6 characters English Message

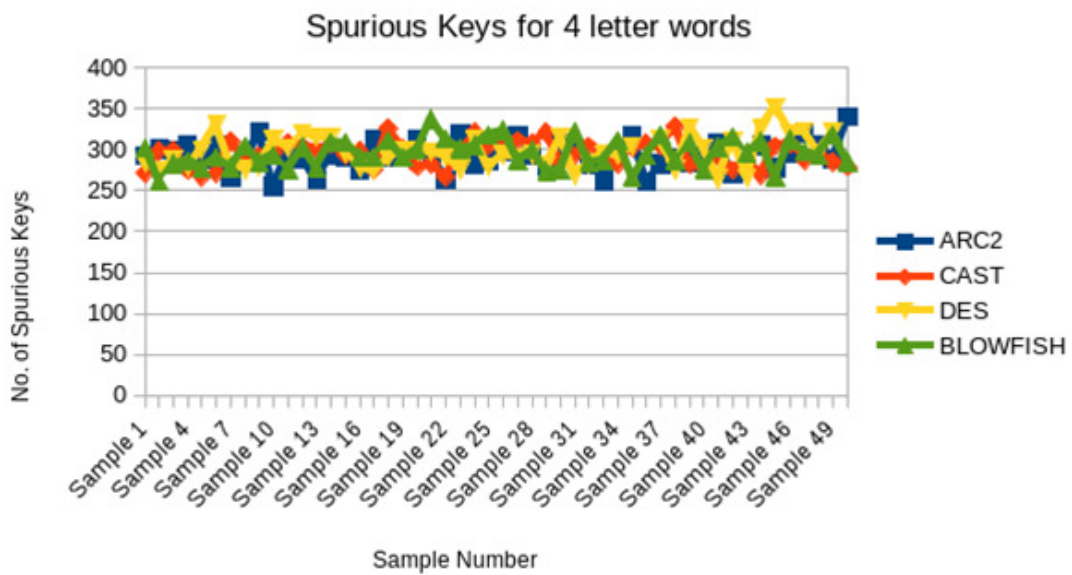


Figure 5. Spurious key analysis for 4 characters English Message

Comparative analysis of the entire evaluation is presented in the Table 2. The number of spurious keys is approximately same for each algorithm due to the fact that the message text is also same. The key length and message lengths are bounded by the block cipher principles based on which the spurious keys exhibit similar characteristic on all ciphers considered in this evaluation.

Table 2. The average number of spurious keys for English text

| | ARC2 | CAST | BLOWFISH | DES |
|----------|-------------|-------------|-----------------|------------|
| 4 LETTER | 292.86 | 294.18 | 295.82 | 296.74 |
| 6 LETTER | 294.56 | 296.22 | 298.2 | 293.2 |
| 8 LETTER | 291.78 | 290.62 | 287.5 | 293.82 |

4. PERFORMANCE ANALYSIS OF BLOCK CIPHERS FOR DEVANAGARI SCRIPT

The spurious key analysis is performed for Devanagari Script also. Devanagari is a south and central Asian language described by Unicode along with Bengali, Gurumukhi, Gujarati, Oriya, Tamil, Telugu, Kannada and Malayalam scripts. This script is represented with U+0900 to U+097F. The characters that are likely to be present in a message are considered for the analysis, i.e. 'ॐ ँ ॐ ँ ॐ ँः ओ अ आ इ ई उ ऊ ऋ ल्रँ ऐ ए ऐ औ ओ औ क ख ग घ ङ च छ ज झ ञ ट ठ ड ढ ण त थ द ध न न प फ ब भ म य र र लळ व श ष स \ह \$ \$ \$ \$ s \$ा \$ि \$ी \$ु \$ू \$ृ \$ॄ \$ँ \$ें \$े \$ै \$ॉ \$ो \$ो \$ो \$ौ \$् \$ॢ \$ॣ \$। \$॥ क ख ग ज ड ढ फ य ऋ लृ ऌ ॡ । ॥ ०१२३४५६७ ८९ . ' अँ ॐ ॐ ॐ ॐ ॐ ॐ गृ ज्ञ ? इ ब्र '. The \$ symbol may be ignored in the above character set. Unlike English language analysis, the numbers and some special characters are also considered in this analysis.

Table 3 presents some sample spurious keys for selected plaintexts with key as '12345678'. The resultant ciphertext and the decrypted text-like-text are also given in the table.

Table 3. Sample spurious keys for Devanagari script

| Algorithm | Plaintext | Ciphertext | Text like Message | Spurious Key |
|-----------|-----------|---------------|-------------------|--------------|
| DES | पक्षाघात | ॐ ॐ o&ॐ Ti | मंडुइळ्ळ | 0000202g |
| ARC2 | अमीर | hॐ ॐ mXW~\ | तपछधइॐिँ | 00000043 |
| CAST | अखबार | ॐ ॐ #ॐ #ॐ ॐ ॐ | ०ऐबऐँकऑक | 00000588 |
| Blowfish | तर्जुमान | ॐ 'gॐ ॐ @ॐ ॐ | ढककबयगिड | 00000106 |

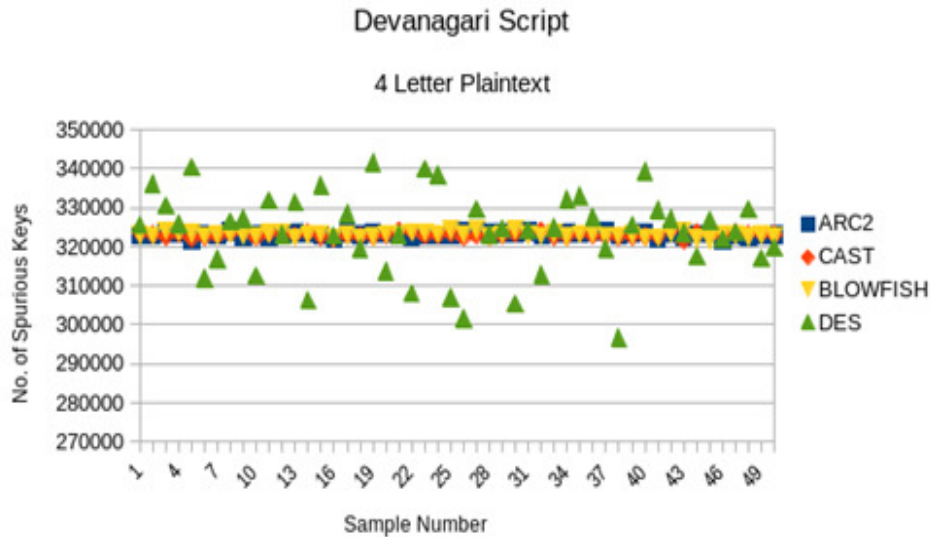


Figure 6. Spurious Keys for 4 Character Devanagari Plain-text

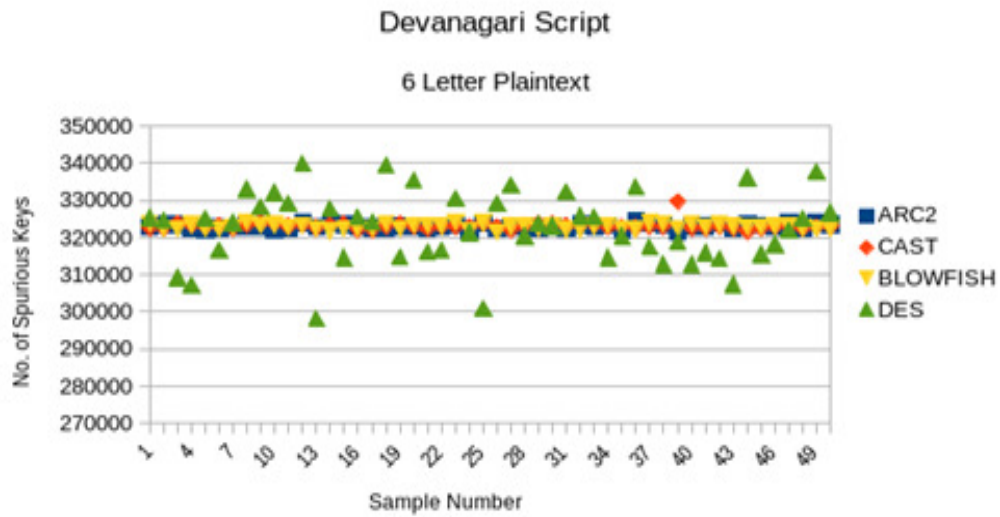


Figure 7. Spurious Keys for 6 Character Devanagari Plain-text

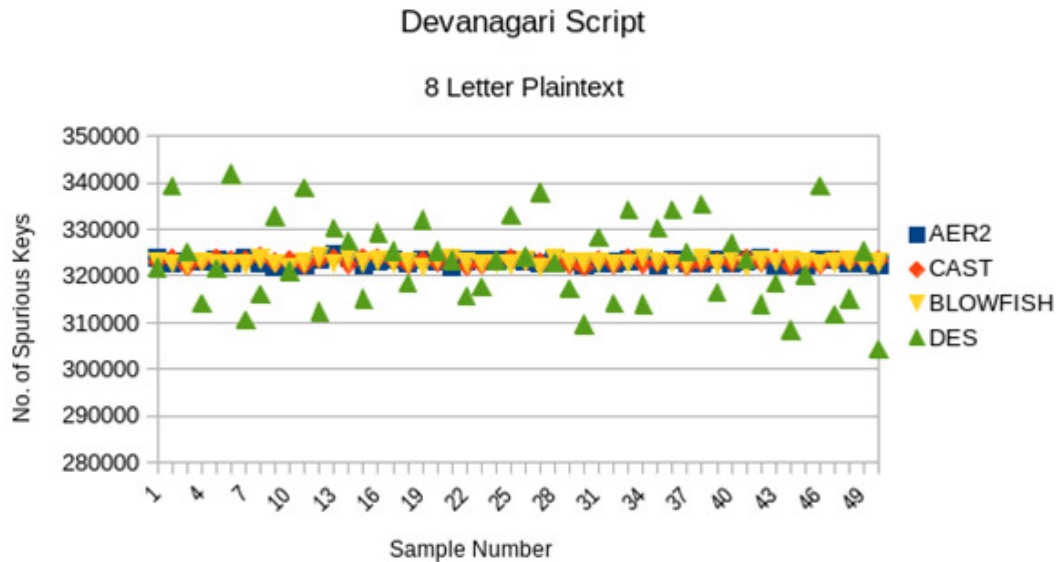


Figure 8. Spurious Keys for 8 Character Devanagari Plain-text

The average number of spurious keys for Devanagari script is summarized in table below. It can be observed that all the four block ciphers behave in similar manner for a given set of plaintexts.

Table 4. The average number of spurious keys for English text

| | ARC2 | CAST | BLOWFISH | DES |
|----------|-----------|-----------|-----------|-----------|
| 4 LETTER | 323231.56 | 323051.72 | 323211.2 | 323007.04 |
| 6 LETTER | 323064.44 | 323172.56 | 322964.58 | 323227.28 |
| 8 LETTER | 323113.74 | 323111.14 | 323001.86 | 323122.3 |

5. CONCLUSIONS

Strength of the cryptosystem is generally identified with the evaluation of weaknesses in terms of algorithm and key. This approach is adopted by many researchers. However, the strength of the cryptosystem should also be identified with the associated strong parameters. An attempt is made in this paper to evaluate the strength of the cryptosystem with the help of spurious keys approach which is termed as strength because there is an inherent confusion arises out of the nature of spurious keys. During decryption, spurious key results in text-like message from the given ciphertext. In an ideal scenario, if every key in the key space acts like a spurious key then the cryptosystem will provide maximum strength which is not possible in real world.

In the present work, a sub set of key space is considered for evaluation of number of spurious keys in that range. Four block ciphers viz ARC2, Blowfish, CAST and DES are considered for this evaluation. Message space is assumed to be alphabets of the respective script only. In this work, we considered English and Devanagari scripts. We evaluated the possible number of spurious keys associated with the sub key space. A comparison is made in terms of varying text size.

All four block ciphers of the present work posses similar characteristics for the varying text size within the limit of the block length. Devanagari script is found with approximately 323000 spurious keys whereas English messages are found with approximately 294 spurious keys in the sub key space of 10^8 keys. Due to the large number of spurious keys associated with Devanagari script, it is difficult to identify the exact key of the cryptosystem, which is considered to be the strength of the system associated with the message texts of Devanagari.

It is necessary to explore the impact of numerals and special characters in the message text on the possible number of spurious keys. The evaluation is in progress. Similarly, evaluation on other block ciphers and varying key length is considered as a future task. The impact of script on cryptosystem is another factor to be considered for evaluation as a future task.

REFERENCES

- [1] C. E. Shannon, "Communication theory of secrecy systems", Bell Syst. Tech. J., vol. 28, pp. 656-715, Oct. 1949.
- [2] Goldwasser Shafi and Silvio Micali, "Probabilistic encryption." Journal of computer and system sciences 28.2 (1984): 270-299.
- [3] Petitcolas et al, "Information hiding-a survey." Proceedings of the IEEE 87.7 (1999): 1062-1078.
- [4] Moulin Pierre and Joseph O'Sullivan, "Information-theoretic analysis of information hiding." Information Theory, IEEE Transactions on 49.3 (2003): 563-593.
- [5] Diffie Whitfield and Martin E. Hellman, "Special feature exhaustive cryptanalysis of the NBS data encryption standard." Computer 6 (1977): 74-84.
- [6] Langford Susan K and Martin E. Hellman, "Differential-linear cryptanalysis." Advances in Cryptology—CRYPTO'94. Springer Berlin Heidelberg, 1994.
- [7] Nyberg Kaisa, "Perfect nonlinear S-boxes." Advances in Cryptology—EUROCRYPT'91. Springer Berlin Heidelberg, 1991.

- [8] Courtois Nicolas T and Josef Pieprzyk, "Cryptanalysis of block ciphers with overdefined systems of equations." *Advances in Cryptology—ASIACRYPT 2002*. Springer Berlin Heidelberg, 2002. 267-287.
- [9] Kaufman Lori M, "Data security in the world of cloud computing." *Security & Privacy, IEEE* 7.4 (2009): 61-64.
- [10] Leach John, "Improving user security behaviour." *Computers & Security* 22.8 (2003): 685-692.
- [11] Rajesh et al, "Genetic algorithmic approach for dynamic request processing in agent cloud platform." *Advance Computing Conference (IACC), 2015 IEEE International IEEE*, 2015.
- [12] Biham Eli and Adi Shamir, "Differential cryptanalysis of DES-like cryptosystems." *Journal of CRYPTOLOGY* 4.1 (1991): 3-72.
- [13] Biham Eli and Adi Shamir, "Differential cryptanalysis of Feal and N-hash." *Advances in Cryptology—EUROCRYPT'91*. Springer Berlin Heidelberg, 1991.
- [14] Rajput Sudheesh K and Naveen K. Nishchal, "Known-plaintext attack on encryption domain independent optical asymmetric cryptosystem." *Optics Communications* 309 (2013): 231-235.
- [15] Khovratovich et al, "Narrow-Bicliques: cryptanalysis of full IDEA." *Advances in Cryptology—EUROCRYPT 2012*. Springer Berlin Heidelberg, 2012. 392-410.
- [16] Biryukov Alex and Johann Großschädl. "Cryptanalysis of the full AES using GPU-like special-purpose hardware." *Fundamenta Informaticae* 114.3-4 (2012): 221-237.
- [17] Jean et al, "Improved cryptanalysis of AES-like permutations." *Journal of Cryptology* 27.4 (2014): 772-798.
- [18] Matsui Mitsuru, "Linear cryptanalysis method for DES cipher." *Advances in Cryptology—EUROCRYPT'93*. Springer Berlin Heidelberg, 1994.
- [19] Liu Yu et al, "Linear cryptanalysis of reduced-round SPECK." *Information Processing Letters* 116.3 (2016): 259-266.
- [20] Yu Xiao-Li et al, "Zero-Correlation Linear Cryptanalysis of Reduced-Round SIMON." *Journal of Computer Science and Technology* 30.6 (2015): 1358-1369.
- [21] Leurent Gaëtan, *Differential and Linear Cryptanalysis of ARX with Partitioning--Application to FEAL and Chaskey*. Cryptology ePrint Archive, Report 2015/968, 2015.
- [22] Biham et al, "Linear cryptanalysis of reduced round Serpent." *Fast Software Encryption*. Springer Berlin Heidelberg, 2002.
- [23] Cho Joo Yeon, "Linear cryptanalysis of reduced-round PRESENT." *Topics in Cryptology-CT-RSA 2010*. Springer Berlin Heidelberg, 2010. 302-317.
- [24] Mansoori S. Davood and H. Khaleghi Bizaki, "On the vulnerability of simplified AES algorithm against linear cryptanalysis." *Int. J. Comp. Sci. Network Security* 7.7 (2007): 257-263.
- [25] Sereshgi Faghihi et al, "Biclique cryptanalysis of MIBS-80 and PRESENT-80 block ciphers." *Security and Communication Networks* 9.1 (2016): 27-33.
- [26] Bogdanov et al, "Biclique cryptanalysis of the full AES." *Advances in Cryptology—ASIACRYPT 2011*. Springer Berlin Heidelberg, 2011. 344-371.

- [27] Schneier Bruce, "Description of a new variable-length key, 64-bit block cipher (Blowfish)." *Fast Software Encryption*. Springer Berlin Heidelberg, 1994.
- [28] Lenstra et al, "Selecting cryptographic key sizes." *Journal of cryptology* 14.4 (2001): 255-293.
- [29] Chang et al, "Biometrics-based cryptographic key generation." *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*. Vol. 3. IEEE, 2004.
- [30] Monroe et al, "Cryptographic key generation from voice." *Security and Privacy, 2001. S&P 2001. Proceedings. 2001 IEEE Symposium on*. IEEE, 2001.
- [31] Teoh et al, "Personalised cryptographic key generation based on FaceHashing." *Computers & Security* 23.7 (2004): 606-614.
- [32] Chen B and Vinod Chandran, "Biometric based cryptographic key generation from faces." *Digital Image Computing Techniques and Applications, 9th Biennial Conference of the Australian Pattern Recognition Society on*. IEEE, 2007.
- [33] Freire et al, "Cryptographic key generation using handwritten signature." *Defense and Security Symposium. International Society for Optics and Photonics*, 2006.
- [34] Mittal Mohit, "Performance Evaluation of Cryptographic Algorithms." *International Journal of Computer Applications (0975–8887) Volume* (2012).
- [35] Mandal Pratap Chandra, "Evaluation of performance of the Symmetric Key Algorithms: DES, 3DES, AES and Blowfish." *Journal of Global Research in Computer Science* 3.8 (2012): 67-70.
- [36] Black et al, "Encryption-scheme security in the presence of key-dependent messages." *Selected Areas in Cryptography*. Vol. 2595. 2002.
- [37] Halevi Shai and Hugo Krawczyk. "Security under key-dependent inputs." *Proceedings of the 14th ACM conference on Computer and communications security*. ACM, 2007.
- [38] Hofheinz Dennis. "Circular Chosen-ciphertext Security with Compact ciphertexts." *Eurocrypt*. 2013.
- [39] Chiasson Sonia, and Paul C. van Oorschot. "Quantifying the security advantage of password expiration policies." *Designs, Codes and Cryptography* (2015): 1-8.
- [40] Biryukov et al, "Key recovery attacks of practical complexity on AES-256 variants with up to 10 rounds." *Advances in Cryptology–EUROCRYPT 2010*. Springer Berlin Heidelberg, 2010. 299-319.
- [41] Standaert et al, "Cryptanalysis of block ciphers: A survey." *UCL Crypto Group* (2003).
- [42] Hussain Iqtadar and Tariq Shah, "Literature survey on nonlinear components and chaotic nonlinear components of block ciphers." *Nonlinear Dynamics* 74.4 (2013): 869-904.
- [43] Braun et al, "Long term confidentiality: a survey." *Designs, Codes and Cryptography* 71.3 (2014): 459-478.
- [44] Alléaume et al, "Using quantum key distribution for cryptographic purposes: A survey." *Theoretical Computer Science* 560 (2014): 62-81.
- [45] Kong et al, "A comprehensive survey of modern symmetric cryptographic solutions for resource constrained environments." *Journal of Network and Computer Applications* 49 (2015): 15-50.
- [46] Korner, "Compressing inconsistent data." *Information Theory, IEEE Transactions on* 40.3 (1994): 706-715.

- [47] Witsenhausen Hans S, "The zero-error side information problem and chromatic numbers (corresp.)." Information Theory, IEEE Transactions on 22.5 (1976): 592-593.
- [48] C. E. Shannon, "The zero-error capacity of a noisy channel," IRE Trans. Inform. Theory, vol. IT-2, pp. 8-19, 1956.

AUTHORS

Mr. Mekala Ramarao is a full time research scholar at JNTUH College of Engineering Hyderabad. He is a life time member of ISTE. His areas of interest include language based security, information theoretic security, embedded security and block chain technology. He is awarded Junior Research Fellowship (JRF) from University Grants Commission (UGC) under Rajiv Gandhi National Fellowship (RGNF) scheme.



Dr. L. Pratap Reddy received, B.E. degree from Andhra University (INDIA) in Electronics and Communication Engineering in 1985, the M.Tech. degree in Electronic Instrumentation from Regional Engineering College (WARANGAL) in 1988 and the Ph.D. degree from Jawaharlal Nehru Technological University (HYDERABAD) in 2001. From 1988 to 1990 he was lecturer in ECE Department of Bangalore Institute of Technology (BANGALORE), from 1991 to 2005 he was faculty member in JNTU College of Engineering (KAKINADA). Since 2006 he is with Department of Electronics and Communication Engineering at JNTU, Hyderabad. His current activity in research and development includes, apart from telecommunication engineering subjects, Pattern Recognition, Information Security, Embedded Systems and Linguistic processing of languages. He published 75 technical papers, articles and reports. He is active member in professional bodies like IETE, ISTE, IE, and CSI. At present he is Working Chairman of SWECHA and Treasurer of Free Software Movement of India.



BHVS Narayana Murthy, is the Director, Research Centre Imarat (RCI), a Defence R&D Organization (DRDO) laboratory. He is the fellow of INAE and IETE. He is a senior member of IEEE and Life Member of AeSI, CSI. His areas of interest include embedded systems, System on Chip, Re-configurable Intelligent systems, Sensor networks and Machine learning.



Maruti Sairam V Annaluru, is working as scientist in Research Centre Imarat(RCI), a Defence R&D Organization (DRDO) laboratory. He is life member of CSI. He is working on the design and development of Real-time Embedded avionics systems software. His areas of interest includes Safe programming, Model based software development, Information security in embedded systems.



INTENTIONAL BLANK

USABILITY TESTING OF FITNESS MOBILE APPLICATION: METHODOLOGY AND QUANTITATIVE RESULTS

Ryan Alturki and Valerie Gay

School of Electrical and Data Engineering,
University of Technology Sydney, Sydney City, Australia

ABSTRACT

Obesity is a major health problem around the world. Saudi Arabia is a nation where obesity is increasing at an alarming rate. Mobile apps could help obese individuals but they need to be usable and personalized to be adopted by those users. This paper aims at testing the usability of a fitness mobile app "Twazon", an app in Arabic language. This paper presents an extensive literature review on the attributes that improve the usability of fitness apps. Then, it explains our methodology and our set up of a trial to test the usability of Twazon app that is popular in Saudi Arabia. The usability attributes tested are effectiveness, efficiency, satisfaction, memorability, errors, learnability and cognitive load. The trial is done in collaboration with participants from the Armed Forces Hospitals - Taif Region in Saudi Arabia. The results highlight that the app failed to meet with the usability attributes.

KEYWORDS

Usability, Mobile Application, Obesity, User Experience

1. INTRODUCTION

Obesity is defined as an excessive storage of energy in the form of fat [1]. According to the facts provided by World Health Organization (WHO) Media Centre, 13% of world's adult population is considered obese, and 39% of the adult population is believed to be overweight. The prevalence of obesity around the world has doubled between 1980 and 2014 [2]. Saudi Arabia is one such country where obesity is increasing at an alarming rate. The study by Coronary Artery Disease in Saudis (CADISS) in 2005 found that 35.5% of people in the country are obese which means every third person in the country is affected. A National Nutrition Survey of 2007 mentioned that obesity is a significant concern because the prevalence of obesity in the men (14%) and women (23.6%) in Saudi Arabia [3]. Overweight and obesity are considered as major risk factors for various chronic diseases such as cancer, cardiovascular diseases and diabetes [4-8].

The health problems and diseases that result from obesity have encouraged a lot of researchers to discuss how the condition can be overcome or prevented [9-14]. Most of the research work states that obesity can be overcome by increasing physical activity and changing eating behaviour. However, it is sometimes very difficult to motivate obese individuals to change their lifestyle and become involved in physical activity. Research has shown that the most effective behaviour change related to fitness and health occurs through behaviour interventions [15-18]. Mobile technology such as mobile applications (apps) have been found to be a very useful intervention

tool for increasing physical activity because through their unique features these apps motivate individuals to achieve their fitness goals [19-22].

Fitness apps are becoming increasingly popular both around the world and in Saudi Arabia. Smartphones and their apps have seen an exponential growth in their usage in Saudi Arabia in recent times. Researchers ranked the country third overall in terms of global smartphone usage penetration [23]. In 2016, smartphone users in Saudi Arabia are estimated to be near 15.9 million and this figure is estimated to increase to 19.1 million by 2019 [24]. Because of cultural restrictions many people, but especially women, find it easier to interact publically and socially in a virtual environment through mobile apps on smartphones. The increasing ubiquity of smartphone technology provides an opportunity to develop an Arabic app to help fight obesity. Sometimes apps development can cost millions of dollars but most of the apps fail miserably [25]. Of all the branded apps, 80% are downloaded less than one thousand times and only 1% has been downloaded one million times or more. After downloading, 25% of mobile apps are never used again [26]. Ample market research suggests that the main reason of failure of mobile apps is the lack of usability [25, 27-28]. The usability of mobile apps enhances user experience (UX) and can play a significant part in the success of the mobile apps [29-32].

The population of obese and overweight individuals in Saudi Arabia is increasing at an alarming rate. Therefore, the increasing use of health and fitness apps in Saudi Arabia is an opportunity to introduce a technological solution which involves developing an app will be popular among obese individuals. There are many Arabic health and fitness apps available but to our knowledge none of these apps have been built with the purpose of enhancing the usability of the app to motivate people to lose weight by considering usability attributes and factors. Moreover, there is no research on fitness and health apps that outlines guidelines for usability. Moreover, to the best of our knowledge, there is no Arabic app that uses any specific features that enhances UX for obese individuals. This leads to our research problem:

- How to improve mobile fitness apps usability to help obese users to reach their health and fitness goals?

In order to solve the above issue, we will start by testing the usability of a popular fitness mobile app that is targeted Saudi users. Then we will develop a new usability guideline that will be specifically designed for fitness apps that help obese users to lose weight. Based on this guideline, we will develop a new fitness app. Our primary focus will be users in Saudi Arabia because the country has a high percentage of obesity rate among its citizens.

This paper presents our method and results of the usability of “Twazon”, an Arabic-language fitness mobile apps. Seven usability attributes: effectiveness, efficiency, satisfaction, memorability, errors, learnability and cognitive load were tested collaboration with the Armed Forces Hospitals in the Taif Region of Saudi Arabia, which provided the participants. All participants are people suffering from obesity and are motivated to lose weight.

2. RELATED WORK

2.1. Fitness Apps in Saudi Arabia

Alnasser et al. examined 65 Arabic fitness apps to determine the level of adherence for each app to the 13 evidence-informed practices [33]. The Centers for Disease Control and Prevention, National Institutes of Health, the Food and Drug Administration and the US Department of Agriculture determined these 13 evidence-informed practices [34]: 1-BMI is determined and explained; 2-fruits and vegetables are recommended and tracked for daily servings; 3-Physical

activities are recommended for daily use; 4-recommendations for drinking water and tracking the daily consumption; 5-recording and tracking the daily consumption of food; 6- a calorie tracker is provided for maintaining calorie balance; 7-advising goal-setting to lose 1 to 2 lb per week; 8-portion control information is provided; 9-Advising users about ways to read and understand nutrition labels; 10- a weight-tracking feature should be provided; 11-physical activities are tracked for daily use; 12-recommending and providing a tool for planning meals; 13-providing a social network among users or allowing users to share via popular social networks, for example Facebook, Twitter, Instagram or Snapchat.

The result of this study stated that there is no app that has more than six evidence-informed practices and only nine apps had between four to six. Therefore, it is clear that there is no Arabic fitness app which successfully adheres to all 13 practices and there is an essential need to address this issue.

2.2. App Selection Process

All of the apps in Google Play or Apple Store are not free. Fitness apps can be divided into three levels:

- At Level 1, apps are free to download but they do not have all the features. The user needs to subscribe and make payments to access extra features;
- At Level 2, the apps are not completely free. The user must pay to download the apps;
- In Level 3 the apps are completely free to be downloaded.

We selected a fitness app to examine its usability and identify how the features in the app affect UX. The app has a high level of adherence to the 13 evidence-informed practices. The app selected is popular in Saudi Arabia and has high ratings on both Google Play and Apple Store so we expect this app to have special features and to be usable. The study of the app will help us determine how the usability in fitness apps can be increased. The app selected is free so that the participants in the usability testing can access them without cost to themselves.

Twazon: We chose Twazon because it is an app developed by academics and it has ten evidence-informed practices out of 13. Another advantage of using the Twazon app is that we can measure the impact of language on the UX. The app is built to make simpler the necessary changes in key diet and exercise behaviour amongst Saudi adults whilst also considering cultural norms. It is also compatible for integration with the Health app on the iPhone. The app is compatible both with Android and iOS operating systems. Twazon has a 4+ rating on both Google Play and Apple Store [35-37].

2.3. Usability

Usability is considered one of the main factors that define the success of a smartphone [38]. Usability can be defined as a multidimensional characteristic of any product. International Standards Organization (ISO) standard 9241-11 gives the meaning of usability as the “extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” [39-40]. This definition of usability has been accepted widely [41]. In 2011 it was replaced by ISO/IEC 25010. This form includes a model of software quality that portrays usability as the degree to which a satisfied user can efficiently and effectively attain certain goals under specific conditions. The term UX is used extensively in contrast to usability. The two terms are used interchangeably. However, UX has a much broader meaning than the term usability [42]. It can say that usability is more concerned

with how easy the product and display features are to use. UX includes the user and the product's complete interaction as well as the thoughts, feelings and perceptions that are the results of this interaction [43].

2.4. Usability Models

2.4.1. ISO Usability model

The ISO98 identified three usability attributes, which are effectiveness (demonstrating the level of accuracy and completeness of goal achievement); efficiency (how well resources were used for the sake of effectiveness); and satisfaction (relief, and positive user interaction whilst operating the software). The ISO98 further outlined those usability factors that needed to be considered. These were: user (the person interacting); goal (or main objective); and the background of use (including users, tasks, tool used, and environment). Each one of these factors affects overall how the software will be designed. Specifically, it affects user interaction with the system [39, 44].



Figure 1. IOS usability model

2.4.2. Nielsen Usability Model

Nielsen was one of the first to identify the attributes of usability. Whilst Nielsen's earlier model had only four attributes, which are Effectiveness, Efficiency, Satisfaction and Learnability [45]. However, he later removed Effectiveness and included both Memorability and Errors in his new model. He identified the following attributes [46]:

- Efficiency: Resources used in completing a task accurately to achieve user goals;
- Learnability: The ease with which the system can be learnt so that the user can start to use it to perform tasks in the minimum amount of time;
- Satisfaction: The product should provide comfort and also give the user a positive attitude towards using it;
- Errors: The error rate of the system should be minimal so that the user makes the least number of errors when using the system. If some errors are made, they should be recovered from easily. Finally, catastrophic errors should be avoided;
- Memorability: One should be able to easily memorise the system to the extent that when even a casual user begins using it after a substantial period of time, they do not have to put effort into learning everything from the beginning.

Nielsen's research defines the term utility as how effectively the system can meet user needs. This is not part of the usability but rather it is an entirely separate system attribute. The product

that has no utility for the user lacks the functions and features required. Such a product therefore has a superfluous utility and will not help the user achieve their goals [47].

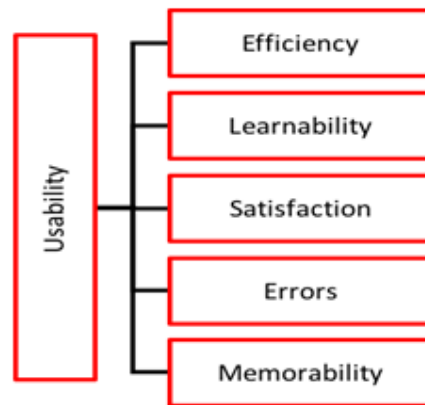


Figure 2. Nielsen usability model

2.4.3. PACMAD Usability Model

This is one of the latest and most frequently used models in recent research on usability. The PACMAD (People At the Centre of Mobile Application Development) model of usability was introduced to overcome issues that have emerged because of the advent of mobile apps. The model aims to include other attributes, which were ignored by other models [47]. Limitations of previous usability models are addressed by PACMAD when they are applied to mobile devices. They included cognitive load in their model because it is likely to have the most significant influence on either the success or failure of an app. The PACMAD model includes attributes for both the ISO and Nielsen models but also incorporates the attributes of both Nielsen's model and ISO standard. The model states that overall usability of a mobile app is affected by 3 factors: task, user, and context of Use. These are recognised by both the ISO and Nielsen. Each of the three factors includes seven attributes. Six are discussed and included in the Nielsen and ISO models. Cognitive load is therefore the new entry and so its inclusion is considered to be PACMAD's main achievement and contribution. Cognitive load is defined as the cognitive processing level that the user requires to use the app. In traditional models, it is assumed that the user is involved in or is performing one task at a time; however, in the context of a mobile device, a user can perform multiple actions whilst using mobile apps. For example, a user might be cooking food whilst he is listening to the stereo.



Figure 3. PACMAD usability model

2.5. Discussion

Most of the research discussed enhancing the usability of mobile apps to help people to be motivated and to meet its desired goals. We believe it has missed one very important aspect, which is that of social and cultural norms. For an app to be usable, it should meet with the social and cultural norms for users - be in the users' native language and considering the social customs. Soroa-Koury and Yang recruited 343 participants in their study to find how traditional views and social customs play a part in the prediction of a user's response to mobile apps. The study demonstrated that social norms predicted perceived ease of use (PEOU) and perceived usefulness (PU) [48]. Many researchers in the past have found that social norms affect human behavior [49-51]. Various studies are available which show that social norms have been used to intervene in undesirable behaviors such as drinking, smoking, and sexual proclivity. Researchers investigating the acceptance of technology have incorporated social norms as important predictors of user behavior when adopting a particular technology [52].

Despite the importance of usability, social and cultural norms on the success of a mobile app, there are very few apps in the field of health and fitness that consider their impacts of usability and cultural and social norms when designing the apps. A research by Alnasser et al. was involved the development of 'Twazon'; about which they claim to consider social and cultural norms of Saudi society [35]. However, they have not mentioned any new feature or attribute that makes the app socially more acceptable or more culturally relevant. Mostly, they have used cultural norms as a reason why women are not physically active. The only thing that makes this app culturally and socially aware is the use of Arabic language. For example, the app does not include any special timetable and diet plan for the month of Ramadan. Ramadan is one of the most important religious periods in the Islamic year during which Muslims are not allowed to eat from dawn until dusk. The app uses the Gregorian calendar instead of the Islamic one. Even though the language of the app is Arabic, it still uses the English numbers. Therefore, there are limits to existing research done to make it culturally and socially more relevant and acceptable. Moreover, the app was designed without considering usability attributes and factors. Usability is a very important feature for the success of mobile apps. Therefore, these open issues are identified:

- There is no fitness app available to Saudi users that was designed with the main objective of helping or assisting obese individuals to reach their fitness goals;
- There is no fitness app available to Saudi users that was designed considering usability attributes and factors;
- There is no fitness app available that was designed considering the impacts of cultural and social norms that targeting individuals suffering from obesity in Saudi Arabia.

This leads to the focus of this research:

- How to improve mobile fitness apps usability to help users reach their health and fitness goals and more specifically it discusses how we set a trial to identify;
- What makes mobile fitness apps usable and useful to be easier to use.

3. METHODOLOGY

3.1. Research Strategy

This paper uses experiment as the main research strategy [53]. The purpose of experimental research design is to assist the researcher to establish a cause-effect relationship with a lot of

credibility. Experiments have a particular nature; they are conducted in a systematic way and under controlled conditions. An artificial situation is formed and events, which go together or have something in common are pulled apart [54]. A widely-used definition for experimental research strategy is where scientists actively influence something to observe the consequences [55]. Experimental research strategy can be categorized into [54]:

- Laboratory experiments: These are carried out in settings that are specially created and the experimenter has the ability to control a variety of extraneous variables;
- Natural experiments: These are referred to as quasi –experiments. These studies are conducted when a natural event or social policy creates situations suitable for the experiment. The investigator has no control over independent variables. These subjects are neither matched in groups nor randomly assigned;
- Field experiments: In these experiments, independent variables are manipulated by the researcher in a field environment.

Moreover, laboratory testing has been used in a lot of usability research. Six techniques were outlined and evaluated for usability testing in a laboratory environment. These techniques facilitated systematic collection of data and identified usability problems experienced by mobile users. According to the result form the research, laboratory testing methodology has huge advantages [56]. Additionally, a laboratory testing was used in usability study, which aimed to discover the impacts of the small screen size of mobile devices upon web browsing and navigation [57]. Moreover, laboratory testing allows us to use evaluation techniques such as ‘think-aloud’ and observation, which cannot be applied through other means. These techniques are not possible in any other research choices such as the field setting [58]. Laboratory testing can reveal a lot about usability; even with the minimum number of participants. According to previous usability guides, 80% of the problems related to usability in any product can be revealed by having four or five participants in the experiment. Similarly, other studies showed that 90% of all usability problems can be detected using ten participants [45, 59-60].

Therefore, in this research, usability testing in a laboratory setting is seen as the most appropriate strategy. A laboratory is a peaceful environment where the users can easily concentrate on the tasks provided to them.

3.2. Usability Metrics Selection

The ISO/IEC 9126-4 makes the recommendation that any usability metric must make reference to effectiveness, efficiency and satisfaction. Attributes such as memorability, errors, cognitive load and learnability are linked to the efficiency and effectiveness of the app. Whilst they each measure the effectiveness and efficiency of these apps, they do so from a specific perspective. If an app has less errors, it means that it is effective because the user can perform more tasks in less time without repeating the tasks with errors. Similarly, if an app has better learnability, it helps the user undertake more tasks accurately so it is more effective. The user therefore becomes more efficient in their completion of these tasks. Any of the above features, improve usability and user satisfaction. This is the reason that usability metrics usually include effectiveness, efficiency and satisfaction as the important features for improving usability.

3.2.1. Usability Metric for Effectiveness

Effectiveness can be measured using the completion rate of tasks. However, another measurement that can be used is the number of mistakes that users make when trying to finish a task.

Effectiveness can therefore be defined as a percentage by utilising the simple equation represented below [61].

$$Effectiveness = \frac{Number\ of\ tasks\ completed\ successfully}{Total\ number\ of\ tasks\ undertaken} \times 100\%$$

3.2.2. Usability Metric for Efficiency

Efficiency is used as a tool to measure the time taken to finish a task. It is usually the time taken by participants to complete a task. Efficiency can be calculated using two methods: Overall Relative Efficiency and Time-Based Efficiency [61].

$$Overall\ Relative\ Efficiency = \frac{\sum_{i=1}^R \sum_{j=1}^n n_{ij} t_{ij}}{\sum_{i=1}^R \sum_{j=1}^n t_{ij}} \times 100\% \quad Time\ Based\ Efficiency = \frac{\sum_{i=1}^R \sum_{j=1}^n \frac{n_{ij}}{t_{ij}}}{NR}$$

Where:

- R: number of users
- N: number of tasks.
- n_{ij} : result for task (i) by user (j). if the task is completed successfully, then $n_{ij} = 1$, otherwise $n_{ij} = 0$.
- t_{ij} = time spent by user “j” to complete task “i”. If the user does not complete the task successfully, then the time will be measured until the moment the user gave up from the task.

3.2.3. Usability Metric for Satisfaction

Users’ satisfaction can be determined through standardized questionnaires that measure satisfaction. These can be dispensed after each task or following the usability testing session. Once the user attempts a task, they are given a questionnaire to measure the difficulty of task and the task level satisfaction. Post-task questions can take various forms: ASQ, Subjective Mental Effort Questionnaire (SMEQ), Single Ease Question (SEQ), Usability Magnitude Estimation (UME) etc. From the above list, we will use SEQ as recommended by Sauro [62]. SEQ has the advantage in that it is brief and simple to answer as well as being easy for the experimenter to conduct and then tally the results. The SEQ in this case is “Overall, how easy or difficult did you find this task?”. This SEQ has a rating scale of 7 points where 1 is very easy and 7 is very difficult. The level of satisfaction is found via a formalized questionnaire for users to gain an overall idea of how easy the app is to use. There are different types of questionnaires available however the choice depends on the budget as well as the degree of significance placed upon the user’s perceived level of satisfaction as a factor of the overall project [63].

3.2.4. Usability Metric for Cognitive Load

Cognitive load has been identified as the measure of mental activity on working memory at any particular instance [64]. To determine the app’s cognitive load, we will use the National Aeronautics and Space Administration (NASA) Task Load Index (TLX) test. NASA-TLX allows the user to evaluate the situation of the workload after the testing is done. It measures the overall task demands by identifying 3 broad scales, which are task, behaviour and subject-related. Each of the scales has factors. The task-related scale includes mental, physical and temporal demands. The behaviour-related scale includes performance and effort. Subject related includes frustration. A user will need to have description for each of the factors as demonstrated below [65]:

- Mental demand: To what extent did you need to perform mental and perceptual activities (such as thinking and calculating)?

- Physical demand: To what extent did you need to perform physical activities (such as pushing and pulling)
- Temporal demand: To what extent did you feel a time pressure while performing tasks?
- Effort: How hard did you have to work hard (mentally and physically) to perform tasks?
- Performance: How satisfied are you with your performance?
- Frustration level: How stressed or annoyed did you feel while performing these tasks?

The NASA-TLX test contains two stages which are weights and ratings. In the weighting procedure, a user will be required to evaluate the influence of each factor regarding a task. There are 15 potential pairs of factors about which a comparison is made. A user will be giving 15 cards and each card contains a pair of the factors and asked to select the most relevant factor regarding the task. Each time the user selects from a pair, the examiner counts it. The scale for a factor for each user can range from 0 to 15. The total comparisons for all factors should equal 15. In the second stage, a user needs to rate each of the factors above in a scale that is divided into 20 equal intervals and each interval equals 5 points with a total of a 100 on the scale. As it is a post-event test, it is an effective one as it captures the thoughts and interaction of the user.

3.2.5. Usability Metric for Learnability

Learnability is the ability of the interface to help the user accomplish tasks on the first attempt [66]. Learnability can therefore be measured through establishing the task performance of users who have not been exposed to that app before. Another way of looking at usability is through perceiving how usability or task performance has improved after repeated trials.

3.2.6. Usability Metric for Errors

Another usability measurement is measuring the amount of errors made by the user when completing a task. Errors are defined as mistakes that are made by the participant when attempting a task. Counting the errors provides excellent diagnostic information and it should be mapped into usability problems [67].

3.2.7. Usability Metric for Memorability

Memorability measures how easy it is to remember how to perform a task on the app after the casual user returns to the app after a certain period of not using it [47]. Memorability has the same tests of efficiency and effectiveness but these are repeated after some period of time in order to determine whether the user has remembered how to perform the same task; and hence whether this has improved the usability.

3.3. Usability Testing Environment

The tests were conducted in a typical usability test environment. Laboratory settings were controlled in order to ensure that there were no external interruptions such as varying lighting conditions or disturbing noises. Test sessions were completed via Apple's wireless AirPlay technology. A MacBook was used for recording. The first step was to install Reflector, which is a wireless streaming and mirroring receiver that converted a laptop into an AirPlay receiver. This allowed the user to mirror their smartphone's screen onto their laptop. It also eliminated the need to have an external camera to record events. Moreover, it also helped to minimize the distraction for the user. The purpose of using this software and technology was to create the friendly and quiet environment that is essential for usability testing [68-69].

Nine participants tested the Twazon app. While they tested it, their mobile screens were recorded through Reflector software. All participants were asked to use the app 3 times. Each time, all participants were asked to perform 14 tasks, which were the same for all users. The time difference between the first and second sessions was one hour. Between the second and the third sessions, there was a one-week interval.

The Armed Force Hospital in the Taif Region of Saudi Arabia provided candidates who suffered from obesity and were motivated to lose weight in order to have a healthier life style. The usability test was divided into five phases:

- Introduction: In the first phase, both participants and the examiner introduced themselves. The purpose of the introduction phase was to establish a comfortable interaction between the examiner and participants.
- Warm-up: In this phase, participants were asked to download the app “Twazon” and to fill out a brief questionnaire that aimed to collect participants’ information such as gender and age.
- Deep focus: During this phase, the examiner gave the users a list of the 14 tasks. The participants used the app with the focus being on what it was doing; how it worked and how the app could be used. The examiner encouraged the participants to think aloud while they were performing the tasks. Moreover, when participants finished a task, they were asked to rate it in an SEQ questionnaire.
- Retrospective: In the penultimate phase, the examiner explained the NASA-TLX questionnaire and asked participants to fill it out.
- Wrap up: In the final phase, the examiner thanked the participants and answered any enquiries.

4. RESULTS

Researchers examined all the videos that were recorded on mobile screens while the participants were performing in the trial. All users who successfully completed a task scored 1 and at the same time we measured how long it took to complete a task. In contrast, users who completed a task in the wrong way or gave up on a task received 0 and the time taken was measured as well. Then the equations for effectiveness, overall relative efficiency and time-based efficiency were applied. Then all errors that participants had made while performing tasks were calculated. Regarding the learnability attribute, we compared participants’ performances in the first session with those of the second. Memorability was then measured by comparing participants’ performances in the second session with those of the third. Both satisfaction and cognitive loads were applied only in the first session as they measured the performances of participants who had not previously been exposed to an app. If these loads had been applied in the second and third sessions, this condition could not have been met. Next, we examined the data from the SEQ questionnaire that was used to measure satisfaction. The rating for each user was calculated and then divided by 14 to determine the average satisfaction value for each user. We then examined the data from the NASA-TLX questionnaire and applied the roles to determine the total user score for the cognitive load [65].

Table 1. Participants' information.

| Users | Gender | Age group | Occupation | Type of phone |
|-------|--------|--------------------|----------------------------|---------------|
| 1 | Male | 35 to 44 | Self employed | iPhone 7 |
| 2 | Male | 25 to 34 | Teacher at high school | iPhone 7 |
| 3 | Female | 25 to 34 | Unemployed | OnePlus 3 |
| 4 | Female | 45 to 54 | Government employee | iPhone 6 S |
| 5 | Female | 25 to 34 | Government employee | HTC 10 |
| 6 | Female | Prefers not to say | Prefers not to say | iPhone 7 |
| 7 | Female | 25 to 34 | Accountant in a company | iPhone 7 Plus |
| 8 | Female | 25 to 34 | Receptionist at a hospital | iPhone 6 S |
| 9 | Female | Prefers not to say | Prefers not to say | iPhone 7 |

Nine participants, seven females and two males, were part of the usability of Twazon app. Their information is presented in Table 1.

4.1. Effectiveness

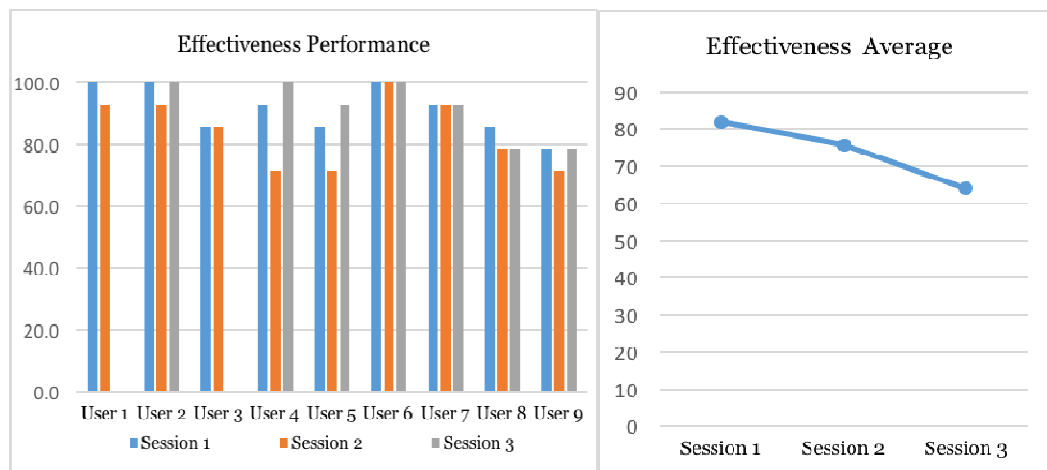


Figure 4. Effectiveness performance

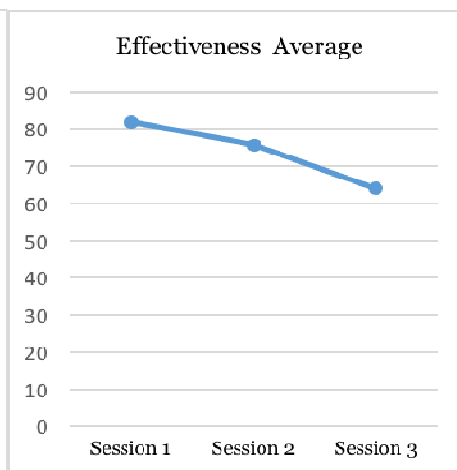


Figure 5. Effectiveness average

Figure 4 describes each user's effectiveness performance over the course of the three sessions. User 6 had the highest percentage of value each time. In session 1, it is 100% and remains constant in session 2 as well as session 3. User 7 showed the same pattern though the value is lower at 92.85%. User 2, user 4, user 5 and user 9 showed positive progress across sessions. However, Users 1 and user 3 had a negative performance because in session 3 they both scored 0%. In addition, User 8's effectiveness performance also slightly decreased. Figure 5 shows the effectiveness performance average, which decreased over each session. In session 1 it was 82%, then it fell to 75.71% and finally in session 3, it reached to 64%.

4.2. Efficiency

4.2.1. Overall Relative Efficiency

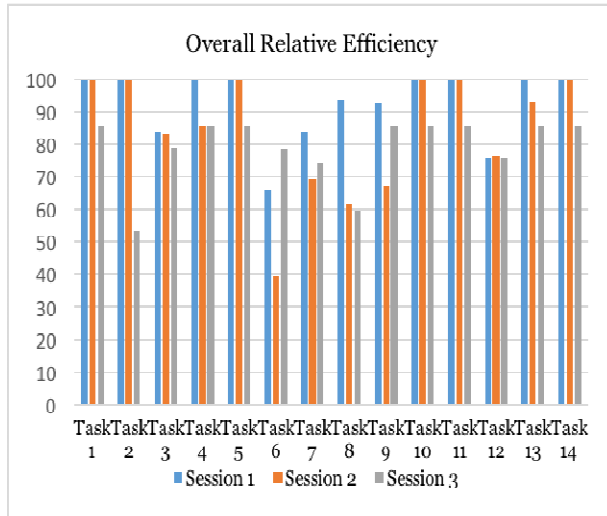


Figure 6. Overall relative efficiency

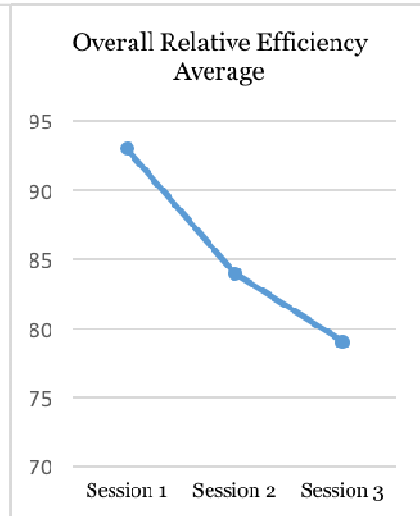


Figure 7. Overall relative efficiency average

Figure 6 demonstrates overall relative efficiency for tasks for the three sessions. In session 1, among the 14 tasks, 8 tasks scored 100% whereas in session 2 and session 3 it was only 6 and 0 respectively. Only in tasks 6 and 12 did the overall relative efficiency percentage improve over each session. However, all the other tasks dramatically decrease between the first and final sessions. Figure 7 shows the overall relative efficiency average, which decreased over each of the three sessions. In session 1 it was 93%, then it fell to 84% and in the final session it reached 79.03%.

4.2.2. Time-Based Efficiency

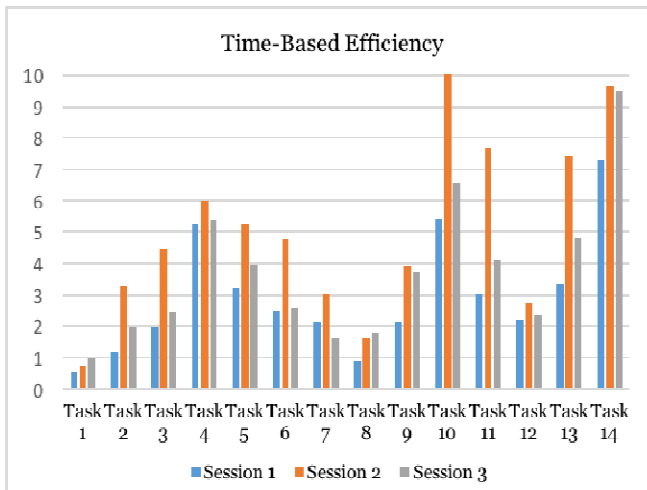


Figure 8. Time-based efficiency

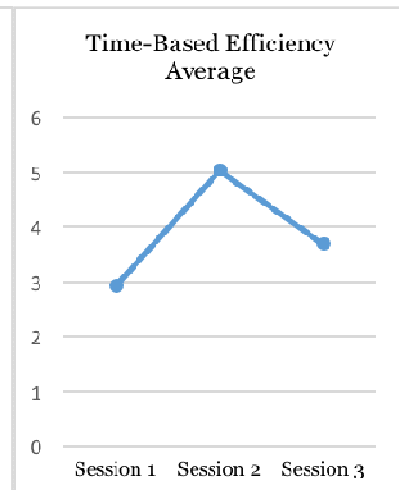


Figure 9. Time-based efficiency average

Figure 8 states time-based efficiency for tasks among the sessions. Task 14 had the highest time-based efficiency score among tasks. In sessions 1, 2 and 3 it was 7.28 goals/sec, 9.66 goals/sec

and 9.48 goals/sec respectively. Task 10 had the second highest time-based efficiency score followed by task 4 and task 13. Interestingly, task 10 reached 10.02 goals/sec in session 2, which was the highest value in all sessions. On the other hand, task 1 got the lowest time-based efficiency followed by task 8 and task 2. Figure 9 shows the time-based efficiency average, which fluctuated across sessions. In session 1 it was 2.93 goals/sec, then it increased to 5.03 goals/sec and finally in the session 3 it decreased to 3.69 goals/sec.

4.3. Satisfaction

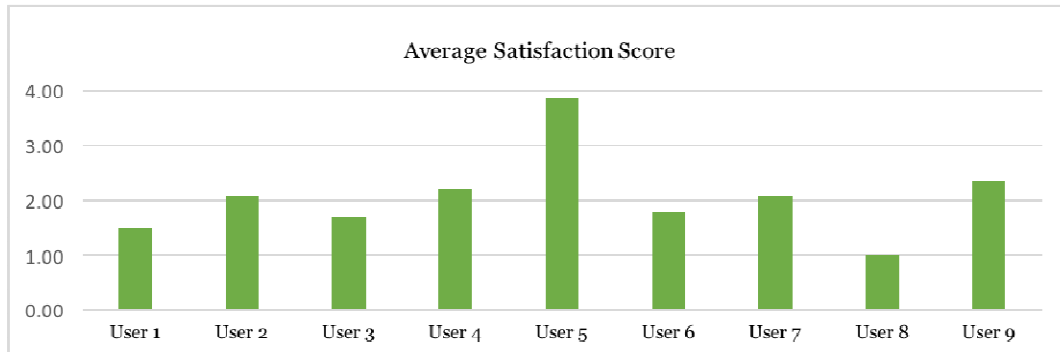


Figure 10. Average satisfaction score

Figure 10 shows each user's average satisfaction score for all tasks. User 5 had highest score at 3.86. User 9 and user 4 scored 2.36 and 2.21 respectively. However, User 8 and user 1 had the lowest score at 1.00 and 1.50 respectively.

4.4. Cognitive Load

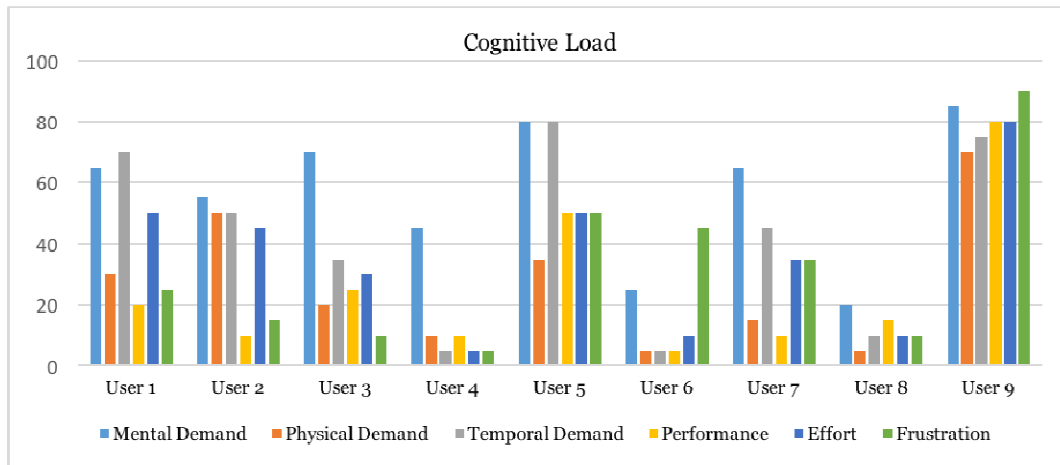


Figure 11. Users' rating for each subscale in cognitive load

Figure 11 shows each user's rating for each subscale in the cognitive load. User 9's cognitive loading is the most consistent. Scores lie between physical demand at (70%) to frustration (90%). However, between user 4 and user 6, the score gap is too high. Mental demand and temporal demand scored the highest value amongst all the subscales. On the other hand, performance and physical demand scored the lowest values.

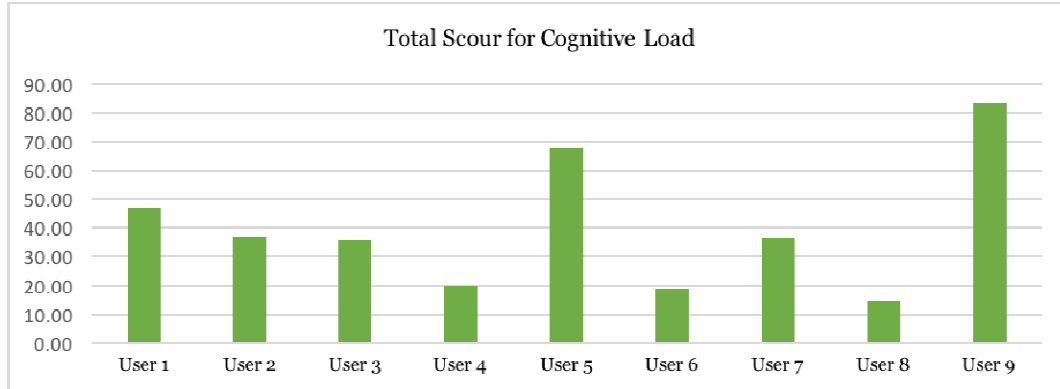


Figure 12. Total score for cognitive load

Figure 12 refers to the total score for cognitive load amongst users. User 9 had the highest value at 83.33%. User 5 and user 1 scored 68% and 47% respectively. However, user 8 had the lowest score at 14.33%.

4.5. Errors

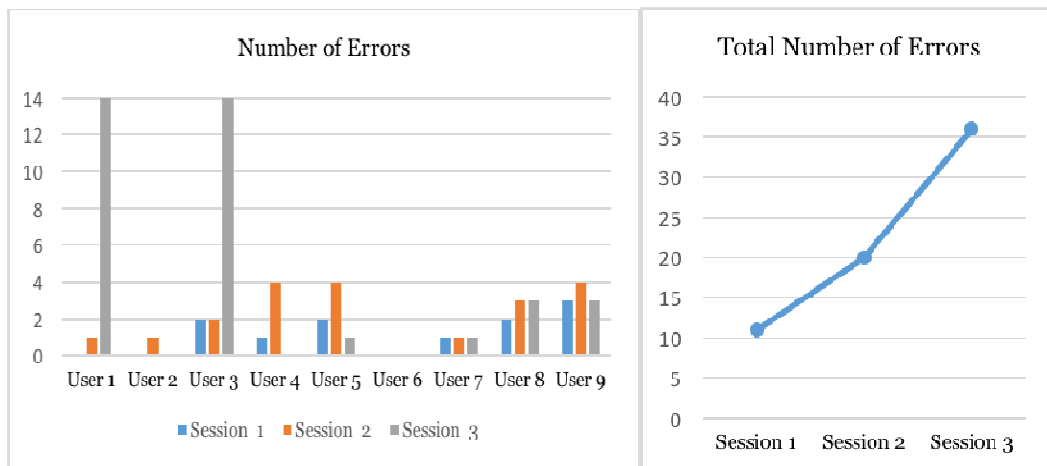


Figure 13. Number of errors

Figure 14. Total number of errors

Figure 13 shows the number of errors made by each user. User 6 is the only user who did not make any errors in all three sessions. User 2 has the second lowest number of errors with just one in session 2. However, User 3 and user 1 had the highest number of errors at 18 and 15 respectively. Figure 14 demonstrates the total number of errors made by all users, which increased over each session. In session 1 it was 11, then it sharply increased to 20 and finally in the third session, it increased to 36.

4.6. Discussion on the Results

One limitation of this study was that users 1 and 3 were not able to participate in the third session as they faced a technical issue with the app. The app did not respond to them when they started performing the first task and after several attempts, they gave up. However, the overall trial for testing the usability of the app succeeded as the level of usability was determined.

Despite the positive increase in the overall score for time-based efficiency between session 1 and session 3, the percentage score for user's effectiveness and overall relative efficiency decreased

over time. Moreover, the number of errors increased from the first session to the second session and did so again from the second to third sessions. As a result of this, the app had a negative association with both learnability and memorability attributes. Furthermore, several participants scored a high percentage in the satisfaction questionnaire, which is negative as a high score means it was more difficult and only one participant rated the whole task as very easy and scored 1 as an average. Besides this, overall cognitive load score was high as the lowest percentage scored by a participant was 14.33%, which means that several participants were not able to perform tasks correctly while doing other activities; for example, speaking to examiners.

The five usability attributes (effectiveness, efficiency, learnability, memorability and errors) did not improve over time. Moreover, both satisfaction and cognitive load scored high percentages because the majority of participants found the app difficult to use. Therefore, the results state that Twazon app has a low level of usability, which is expected due to the fact that it was designed and developed without considering usability attributes and factors.

It is recommended that conducting a qualitative study to determine the reasons and factors that negatively affect the level of usability of the Twazon app. The qualitative study will also consider the importance of social and cultural norms and how they can be applied to improve the usability of the app. A specific usability guideline for fitness mobile apps will then be created, which will help to develop a fitness app that is specially designed for obese individuals in Saudi Arabia.

5. CONCLUSION

The primary purpose of this paper has been to examine the usability for an Arabic fitness mobile app "Twazon". This paper has highlighted the attributes that are considered to be a crucial for improving the usability of fitness mobile apps through presenting an extensive literature review. The paper has presented the methodology and the procedures for testing the Twazon app. Seven usability attributes, (effectiveness, efficiency, satisfaction, memorability, errors, learnability and cognitive load) were tested. The trial for the test was done in collaboration with the Armed Forces Hospitals - Taif Region in Saudi Arabia, which provides the candidates. The result from this trial was that Twazon app failed to meet with the usability attributes and consequently participants found it difficult to use. Future work will include performing a qualitative study for the app to determine how to improve the level of usability and then create usability guidelines for fitness mobile apps. Based on these guidelines, an app that is specifically designed for obese individuals in Saudi Arabia will be developed. Obesity is a major issue for health departments all over the world. Saudi Arabia is a country where the obesity has reached an alarming rate of 35.5% of the population. Better app usability would help keep these individuals motivated to make necessary lifestyle changes.

REFERENCES

- [1] A. P. SIMOPOULOS and T. B. VAN ITALLIE, "Body weight, health, and longevity," *Annals of internal medicine*, vol. 100, pp. 285-295, 1984.
- [2] W. H. Organization. (2016, 2 October). Obesity and overweight. Available: <http://www.who.int/mediacentre/factsheets/fs311/en/>
- [3] O. R. Center. (2016, 10 October). Obesity in Saudi Arabia. Available: <https://www.obesitycenter.edu.sa/pages/patients.aspx?id=258>
- [4] A. Afshin, M. H. Forouzanfar, M. B. Reitsma, P. Sur, K. Estep, A. Lee, et al., "Health Effects of Overweight and Obesity in 195 Countries over 25 Years," *The New England journal of medicine*, vol. 377, pp. 13-27, 2017.
- [5] K. Singer and C. N. Lumeng, "The initiation of metabolic inflammation in childhood obesity," *The Journal of clinical investigation*, vol. 127, pp. 65-73, 2017.

- [6] K. R. Fontaine, D. T. Redden, C. Wang, A. O. Westfall, and D. B. Allison, "Years of life lost due to obesity," *Jama*, vol. 289, pp. 187-193, 2003.
- [7] J. Stevens, J. Cai, E. R. Pamuk, D. F. Williamson, M. J. Thun, and J. L. Wood, "The effect of age on the association between body-mass index and mortality," *New England Journal of Medicine*, vol. 338, pp. 1-7, 1998.
- [8] E. E. Calle, M. J. Thun, J. M. Petrelli, C. Rodriguez, and C. W. Heath Jr, "Body-mass index and mortality in a prospective cohort of US adults," *New England Journal of Medicine*, vol. 341, pp. 1097-1105, 1999.
- [9] C. Summerbell, E. Waters, L. Edmunds, S. Kelly, T. Brown, and K. Campbell, "Interventions for preventing obesity in children (Review)," *Cochrane library*, vol. 3, pp. 1-71, 2005.
- [10] W. Saris, S. Blair, M. Van Baak, S. Eaton, P. Davies, L. Di Pietro, et al., "How much physical activity is enough to prevent unhealthy weight gain? Outcome of the IASO 1st Stock Conference and consensus statement," *Obesity reviews*, vol. 4, pp. 101-114, 2003.
- [11] O. Bar-Or, "Juvenile obesity, physical activity, and lifestyle changes: Cornerstones for prevention and management," *The physician and sportsmedicine*, vol. 28, pp. 51-58, 2000.
- [12] J. L. Anderson, E. M. Antman, S. R. Bailey, E. R. Bates, J. C. Blankenship, D. E. Casey Jr, et al., "AHA Scientific Statement," *Circulation*, vol. 120, pp. 2271-2306, 2009.
- [13] J. O. Hill and H. R. Wyatt, "Role of physical activity in preventing and treating obesity," *Journal of Applied Physiology*, vol. 99, pp. 765-770, 2005.
- [14] J. O. Hill and J. C. Peters, "Environmental contributions to the obesity epidemic," *Science*, vol. 280, pp. 1371-1374, 1998.
- [15] I. Contento, G. I. Balch, Y. L. Bronner, L. Lytle, S. Maloney, C. Olson, et al., "The effectiveness of nutrition education and implications for nutrition education policy, programs, and research: a review of research," *Journal of nutrition education (USA)*, 1995.
- [16] G. D. Foster, A. P. Makris, and B. A. Bailer, "Behavioral treatment of obesity," *The American journal of clinical nutrition*, vol. 82, pp. 230S-235S, 2005.
- [17] T. A. Wadden and A. J. Stunkard, *Handbook of obesity treatment*: Guilford Press, 2002.
- [18] K. D. Brownell, *LEARN program for weight management 2000*: American Health, 2000.
- [19] J. Yang, "Toward physical activity diary: motion recognition using simple acceleration features with mobile phones," in *Proceedings of the 1st international workshop on Interactive multimedia for consumer electronics*, 2009, pp. 1-10.
- [20] T. Denning, A. Andrew, R. Chaudhri, C. Hartung, J. Lester, G. Borriello, et al., "BALANCE: towards a usable pervasive wellness application with accurate activity inference," in *Proceedings of the 10th workshop on Mobile Computing Systems and Applications*, 2009, p. 5.
- [21] S. M. Arteaga, M. Kudeki, A. Woodworth, and S. Kurniawan, "Mobile system to motivate teenagers' physical activity," in *Proceedings of the 9th International Conference on Interaction Design and Children*, 2010, pp. 1-10.
- [22] D. E. Conroy, C.-H. Yang, and J. P. Maher, "Behavior change techniques in top-ranked mobile apps for physical activity," *American journal of preventive medicine*, vol. 46, pp. 649-652, 2014.
- [23] F. Raben and E. Snip, "The MENAP region is developing, but can it keep its promise?," *Research World*, vol. 2014, pp. 6-11, 2014.
- [24] Statista. (2017, 8 March). Number of smartphone users in Saudi Arabia from 2014 to 2021 (in millions)*. Available: <https://www.statista.com/statistics/494616/smartphone-users-in-saudi-arabia/>
- [25] D. LLP. (2012, 2 February). So Many Apps -- So Little To Download. Available: <http://www.mondaq.com/x/192692/IT+internet/So+Many+Apps+So+Little+To+Dow%20nload>
- [26] S. Dredge, "Most branded apps are a flop says Deloitte. But why," ed, 2011.
- [27] M. Bhuiyan, A. Zaman, and M. H. Miraz, "Usability Evaluation of a Mobile Application in Extraordinary Environment for Extraordinary People," *arXiv preprint arXiv:1708.04653*, 2017.
- [28] R. Youens. (2011, 2 February). 7 Habits of Highly Effective Apps. Available: <https://gigaom.com/2011/07/16/7-habits-of-highly-effective-apps/>
- [29] I. Nascimento, W. Silva, A. Lopes, L. Rivero, B. Gadelha, E. Oliveira, et al., "An Empirical Study to Evaluate the Feasibility of a UX and Usability Inspection Technique for Mobile Applications," in *28th International Conference on Software Engineering & Knowledge Engineering*, California, USA, 2016.
- [30] H. Hoehle, R. Aljafari, and V. Venkatesh, "Leveraging Microsoft's mobile usability guidelines: Conceptualizing and developing scales for mobile application usability," *International Journal of Human-Computer Studies*, vol. 89, pp. 35-53, 2016.

- [31] S. Pagoto, K. Schneider, M. Jovic, M. DeBiaise, and D. Mann, "Evidence-based strategies in weight-loss mobile apps," *American journal of preventive medicine*, vol. 45, pp. 576-582, 2013.
- [32] A. C. King, E. B. Hekler, L. A. Grieco, S. J. Winter, J. L. Sheats, M. P. Buman, et al., "Harnessing different motivational frames via mobile phones to promote daily physical activity and reduce sedentary behavior in aging adults," *PloS one*, vol. 8, p. e62613, 2013.
- [33] A. A. Alnasser, R. E. Amalraj, A. Sathiaselan, A. S. Al-Khalifa, and D. Marais, "Do Arabic weight-loss apps adhere to evidence-informed practices?," *Translational behavioral medicine*, vol. 6, pp. 396-402, 2016.
- [34] E. R. Breton, B. F. Fuemmeler, and L. C. Abrams, "Weight loss—there is an app for that! But does it adhere to evidence-informed practices?," *Translational behavioral medicine*, vol. 1, pp. 523-529, 2011.
- [35] A. Alnasser, A. Sathiaselan, A. Al-Khalifa, and D. Marais, "Development of 'Twazon': An Arabic App for Weight Loss," *JMIR research protocols*, vol. 5, p. e76, 2016.
- [36] F. Al-Maarik. (2016, 5 January). Twazon. Available: <https://itunes.apple.com/us/app/twazon/id946772876?mt=8>
- [37] G. Play. (2016, 5 January). Twazon Available: <https://play.google.com/store/apps/details?id=com.twazon.twazon>
- [38] R. Baharuddin, D. Singh, and R. Razali, "Usability dimensions for mobile applications—A review," *Res. J. Appl. Sci. Eng. Technol*, vol. 5, pp. 2225-2231, 2013.
- [39] W. ISO, "9241-11. Ergonomic requirements for office work with visual display terminals (VDTs)," *The international organization for standardization*, vol. 45, 1998.
- [40] S. Ben and C. Plaisant, "Designing the user interface 4 th edition," ed: Pearson Addison Wesley, USA, 2005.
- [41] E. Folmer and J. Bosch, "Architecting for usability: a survey," *Journal of systems and software*, vol. 70, pp. 61-78, 2004.
- [42] D. Saffer, "Designing for Interaction: Creating Smart Applications and Clever Devices," *New Riders Press*, < <http://www.designingforinteraction.com>, vol. 2, p. 2.1, 2007.
- [43] W. Albert and T. Tullis, *Measuring the user experience: collecting, analyzing, and presenting usability metrics*: Newnes, 2013.
- [44] C. Stary and C. Stephanidis, *User-Centered Interaction Paradigms for Universal Access in the Information Society: 8th ERCIM Workshop on User Interfaces for All*, Vienna, Austria, June 28-29, 2004. Revised Selected Papers vol. 3196: Springer Science & Business Media, 2004.
- [45] J. Nielsen, *Usability engineering*: Elsevier, 1994.
- [46] J. NIELSEN. (2012, 18 November). Usability 101: Introduction to Usability. Available: <https://www.nngroup.com/articles/usability-101-introduction-to-usability/>
- [47] R. Harrison, D. Flood, and D. Duce, "Usability of mobile applications: literature review and rationale for a new usability model," *Journal of Interaction Science*, vol. 1, pp. 1-16, 2013.
- [48] S. Soroa-Koury and K. C. Yang, "The Effects of Social Norms on Consumers' Responses to Mobile Advertising," in *Proceedings of the 2009 Academy of Marketing Science (AMS) Annual Conference*, 2015, pp. 162-166.
- [49] M. K. Lapinski and R. N. Rimal, "An explication of social norms," *Communication Theory*, vol. 15, pp. 127-147, 2005.
- [50] G. L. Cohen and D. K. Sherman, "The psychology of change: Self-affirmation and social psychological intervention," *Annual Review of Psychology*, vol. 65, pp. 333-371, 2014.
- [51] R. N. Rimal, M. K. Lapinski, R. J. Cook, and K. Real, "Moving toward a theory of normative influences: How perceived benefits and similarity moderate the impact of descriptive norms on behaviors," *Journal of health communication*, vol. 10, pp. 433-450, 2005.
- [52] V. Venkatesh and F. D. Davis, "A theoretical extension of the technology acceptance model: Four longitudinal field studies," *Management science*, vol. 46, pp. 186-204, 2000.
- [53] M. Saunders, P. Lewis, and A. Thornhill, "Research methods for business students," Harlow: Prentice Hall, 2009.
- [54] M. Rao and S. Shah. (2002, 12 December). EXPERIMENTATION A RESEARCH METHODOLOGY. Available: <http://www.public.asu.edu/~kroel/www500/EXPERIMENTATION%20Fri.pdf>
- [55] O. Blakstad. (2008, 3 January). Experimental Research. Available: <https://explorable.com/experimental-research>
- [56] E. Beck, M. Christiansen, J. Kjeldskov, N. Kolbe, and J. Stage, "Experimental evaluation of techniques for usability testing of mobile systems in a laboratory setting," 2003.

- [57] A. Parush and N. Yuviler-Gavish, "Web navigation structures in cellular phones: the depth/breadth trade-off issue," *International Journal of Human-Computer Studies*, vol. 60, pp. 753-770, 2004.
- [58] N. Sawhney and C. Schmandt, "Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments," *ACM transactions on Computer-Human interaction (TOCHI)*, vol. 7, pp. 353-383, 2000.
- [59] J. Rubin, "Handbook of usability testing: how to plan, design, and conduct effective tests," Wiley technical communication library Show all parts in this series, 1994.
- [60] J. S. Dumas and J. Redish, *A practical guide to usability testing*: Intellect Books, 1999.
- [61] J. Mifsud. (2015, 3 November). Usability Metrics – A Guide To Quantify The Usability Of Any System. Available: <http://usabilitygeek.com/usability-metrics-a-guide-to-quantify-system-usability/>
- [62] J. Sauro. (2010, 9 December). IF YOU COULD ONLY ASK ONE QUESTION, USE THIS ONE. Available: <https://measuringu.com/single-question/>
- [63] A. Garcia. (2013, 18 October). UX Research | Standardized Usability Questionnaire. Available: <https://chaione.com/blog/ux-research-standardizing-usability-questionnaires/>
- [64] J. P. Tracy and M. J. Albers, "Measuring cognitive load to test the usability of web sites," in *Annual Conference-society for technical communication*, 2006, p. 256.
- [65] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," *Advances in psychology*, vol. 52, pp. 139-183, 1988.
- [66] J. Sauro. (2013, 1 December). HOW TO MEASURE LEARNABILITY. Available: <https://measuringu.com/measure-learnability/>
- [67] J. Sauro. (2011, 4 December). 10 ESSENTIAL USABILITY METRICS. Available: <https://measuringu.com/essential-metrics/>
- [68] C. Walsh. (2015, 6 November). A Guide To Simple And Painless Mobile User Testing. Available: <https://www.smashingmagazine.com/2015/12/simple-and-painless-mobile-user-testing/>
- [69] J. Mifsud. (2016, 3 November). Usability Testing Of Mobile Applications: A Step-By-Step Guide. Available: <http://usabilitygeek.com/usability-testing-mobile-applications/>

AUTHORS

Ryan Alturki is a Teaching Assistant in the department of Information Sciences at Umm Al Qura University, Saudi Arabia. Currently, he is doing a PhD in mobile application usability and its role to motivate people to lose more weight.



Valerie Gay has more than 25 years of research experience in leading research labs in Australia and Europe; Her research focuses on the use of mobile technology to offer more personalised advice and care.



RECOGNITION THE DROPLETS IN GRAY SCALE IMAGES OF DROPWISE CONDENSATION ON PILLARED SURFACES

Helene Martin¹, SolmazBoroomandiBarati^{2,*}, Jean-Charles Pinoli¹,
StephaneValette³ and Yann Gavet¹

¹Ecole Nationale Supérieure des Mines de Saint-Etienne,
LGF UMR CNRS 5307, SAINT-ETIENNE, France

²Univ Lyon, Ecole Nationale Supérieure des Mines de Saint-Etienne,
LGF UMR CNRS 5307, SAINT-ETIENNE, France

³Univ Lyon, Ecole Centrale de Lyon, LTDS UMR CNRS 5513, F-69134,
LYON, France

ABSTRACT

This study deals with developing an image processing algorithm that is able to recognize spherical and ellipsoidal droplets growing on pillared surfaces during heterogenous dropwise condensation. The problem with recognizing droplets on the pillared substrates is that droplets are very similar to the pillars or they cover several pillars at the same time, so characterizing the pillars is very important. On the other hand the droplets are not always perfectly spherical or they are connected and form irregular shapes, in such cases the ability to recognizing and separating connected droplets is another challenging step. The method that is used here consists of three main parts: pillars characterization, droplets categorizing and droplets segmentation. The result of this algorithm will be binarized images that enable to extract the information related to the size and density of droplets needed for studying droplets evolution during time. Also a computer simulation method is proposed to generate ellipsoidal droplets on pillared substrate. The results of this algorithm then are validated by comparing with results from experimental procedure.

KEYWORDS

Dropwise Condensation, Image Segmentation, Textured Surface, Droplets Recognition

1. INTRODUCTION

Dropwise condensation has become an attractive process during last decades thanking to its higher heat transfer coefficient (about 5 to 7 times) compared to filmwise condensation [1]. Generally dropwise condensation includes five main steps: nucleation of initial droplets, growth due to adsorption, growth due to coalescence, nucleation of new droplets, and sliding of very big drops.

Nucleation step can occur homogeneously, when there is no preference between different points to grow the droplets, and heterogeneously otherwise. In the process of dropwise condensation, at first small droplets nucleate all around the surface, then these small droplets start to grow by adsorbing water molecules from humid air. After awhile if two or more droplets become big enough to overlap, they will merge and form a bigger droplet -called daughter droplet- in the mass center of the parents [2]. This phenomenon is called coalescence in literature. Although coalescence is a mass conservative process, the area covered by daughter droplet is lower than the summation of area covered by its parents. This will lead to forming vacant area around daughter droplet, in which new small droplets can nucleate and grow. Both these steps will change number of droplets per unit area (N_t) during time and lead to a temporal distribution [3].

Up to know the process of dropwise condensation was mostly studied on flat surfaces that considered as homogenous process [4,5] or on the surfaces with coatings for heterogeneous nucleation [6]. But since the chemical coatings have harmful effects on environment, heterogeneous dropwise condensation on textured surfaces is now more attractive to scientists. Using textured surfaces also enabled us to control droplets configuration on the surface [7]. The problem with condensation on textured surfaces is extracting process information like the radius and density of droplets, especially when they really are similar to the surface texturations. Also the shape of droplets is not perfectly circular in latest steps.

There are lots of methods for droplets recognition in the gray scale images on flat surface. The most adapted methods to recognize spherical droplets in gray scale images are Hough transform [8] and its improvements, such as the normal-line Hough transform [9] and the coherent circle Hough transform [10] methods. Muthukrishnan and Radha [11] used an algorithms based on edge analysis and generated the images by an edge detection method. This method can also be used for droplets that are not perfectly circular. There are also another types of algorithms that use the gray-level intensities as a drop presence indicator. These algorithms use the gray-level intensities directly like the PIV methods [16] and watershed [17] or indirectly like the appearance-based approaches [18]. But the problem with these algorithms on pillared surfaces is the similarity between droplets and pillars and the overlapping droplets. This is because on pillared surfaces the droplets are canalized and form non-circular connected droplets between the pillars.

This study aims to develop an image processing algorithm for recognizing different sizes of droplets on pillared surfaces. This algorithm is able to recognize pillars from droplets or the connected droplets that are not perfectly circular. On the other hand a method for simulating ellipsoidal droplets is proposed that is more suitable for dropwise condensation on textured surfaces.

2. EXPERIMENTAL APPARATUS

Water droplets are formed on the poly carbonate surface of 1 Cm \times 1 Cm with temperature of around 17°C in contact with humid hot vapor, that is chosen 1°C below the dew point of humid air in order to start nucleation of small droplets. Vapor temperature is set at 30°C with relative humidity of 40% that is maintained by a compressor outside the chamber. The substrate is textured by pillars with diameter of 12.5 μ m and height of 3 μ m that are positioned in the distance of 50 μ m. High resolution CCD camera is used to record nucleation and growth of droplets in time interval of 1s. The images taken by CCD camera then are binarized and used to

extract the information of droplets size and number at each time step to validate the results of simulation. The initial density and size of droplets are $3.7 \times 10^7 \text{ m}^{-2}$ and $6 \times 10^{-6} \pm 1 \times 10^{-6} \text{ m}$ respectively.

3. IMAGE PROCESSING

The aim of this section is binarizing the images of real substrate in order to extract the data related to size and number of droplets at each time step. The schematic diagram of the algorithm used to binarize the gray scale images is presented in figure (1).

The problem with pillared surfaces is that pillars are very similar to droplets because of their spherical shape and also because in gray scale images both pillars and droplets exhibit darker boundaries and brighter centers. So the first step will be pillar characterizing in order to separate them from droplets. In this regard, the first image of the considering set is used because the drop presence can be neglected. Then since the images are taken without shifting, one can consider the same pillars for other images. As pillars are perfectly circular with known radius, the circle Hough transform is the most adapted technique. This method consists of implementing an accumulator array which refers to the circle center position probability image for a given radius. For this purpose, an edge detection operator is firstly applied to the original image to get a binary edge image. Then, for each edge point, a circle with this center and the given radius is incremented to the accumulator array. This circle corresponds to the possible center locations of the circle passing through this edge point. Consequently, the peaks of the accumulator array correspond to the most probable center locations of the circles in the images. This principle is used for a range of radius to limit information loss. In our case, as we know the real size of pillars (between 10 and 20 μm), the radius range considered for the Hough transform is [10 - 30] pixels, according to the camera magnification. It is important to consider a quite narrow range in order to get accurate results and a low computational time. To improve again the algorithm performances, the coherent Hough transform is used.

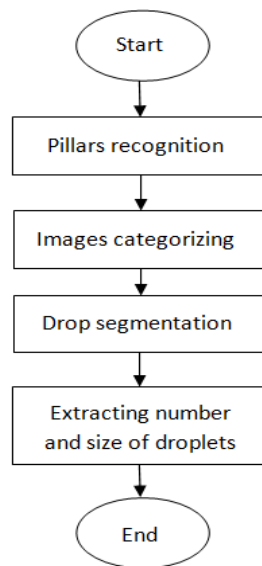


Figure (1): Schematic diagram of image processing algorithm

In the second step the images are divided into images that contain very small, small, medium and big drops. The first two groups refer to the droplets that grow by adsorption. The third group corresponds to the coalescence step and the last one corresponds to the steady state, where the changes in size and number of droplets are negligible. After this categorization, the corresponding image processing algorithm is applied to each group. In this regard, firstly, ten percent of images have been chosen regularly in order to reduce the computation time. Since the time step between images is 1s, 10% of images means selecting one image at each 10 seconds. Then the corresponding gray-tone level histograms of each image are determined as shown in Figure (2). The number of major peaks in each histogram varies as a function of time that represents the different sizes of droplets. Thus, the size ranges of each group are characterized as:

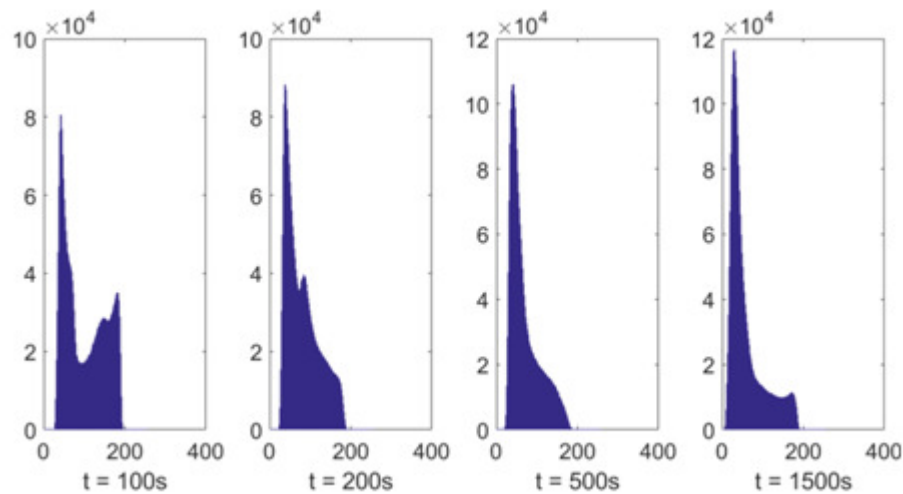


Figure (2): Gray-tone histogram of images taken at $t = 100, 200, 500,$ and $1500s$. The histograms from left to right represent very small, small, medium, and big droplets.

- Very small drops: image with three major peaks representing background, the drop edge and the drop center.
- Small drops: image with two major peaks. In this group the peak corresponding to background is removed because drops tend to cover entirely the sample.
- Medium drops: following the process the drops start to coalesce and their area becomes larger with the same intensity as background. So histogram will show one peak.
- Big drops: finally gray tone histogram shows two major peaks again related to the drop edges and the drop centers. This step corresponds to forming big droplets.

Figure (3) illustrates these four droplets groups. According to this figure the first two groups of droplets are perfectly circular and smaller than the pillars. Also they appear with darker points on the surface that are recognizable. While medium and big droplets tend to have more ellipsoidal shapes with center points very similar to the surface of background. So recognizing these two latest groups needs another additional step to separate droplets from the background.

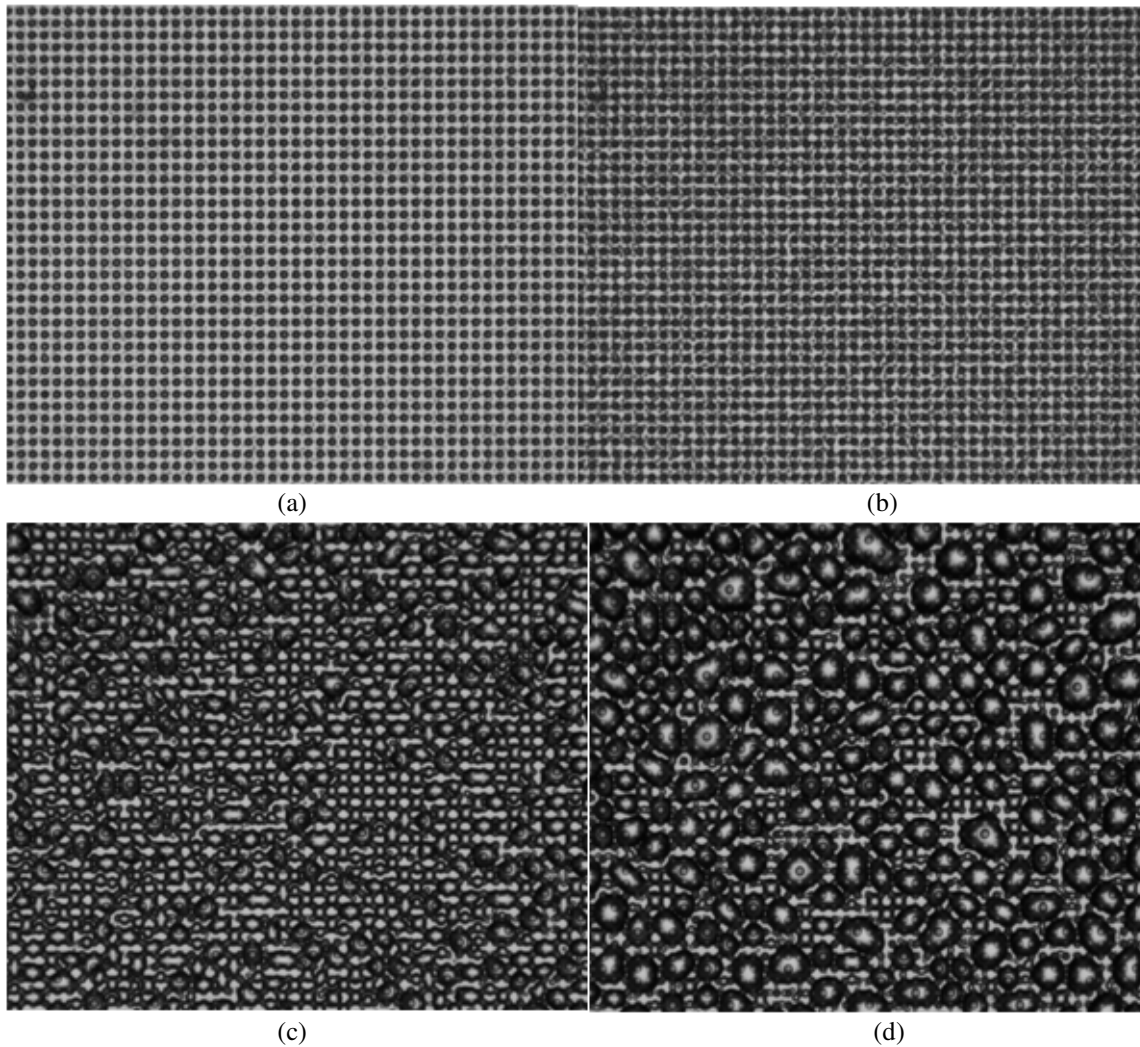


Figure (3): Illustrations of (a) very small droplets, (b) small droplets, (c) medium droplets, and (d) big droplets.

Binarization of each group of droplets then is done with different methods:

3.1 Very small drops:

Very small droplets are appeared in the images as very small dots between the pillars (figure (3,a)). At first, in order to ignoring the pillars, they are removed by means of the method describing in the pillar characterization step. Thus, a gray-level image is obtained with black holes at the pillars places. By comparing the gray tone histograms of very small droplets and the gray tone histogram of image without the droplets, it reveals that the histogram of the first image is less spread than the histogram corresponds to very small droplets. This comparison validates that since the drops are darker than substrate, they modify gray-level intensities of the image towards the vertical axis. The idea is that if we draw the two histograms in the same diagram, the first non-zero point, at which they meet or the non-zero superposition point of these two histograms can be accounted as the thresholding value for gray scale image of very small

droplets. Then, labeling the regions on the thresholded images enable us to have the number of droplets that is equal to the number of regions. Finally mean radius of droplets can be calculated from equation (1) by knowing the total area of droplets (A_{tot}):

$$R_{ave} = \sqrt{\frac{A_{tot}}{N \pi}} \quad \text{Equation (1)}$$

3.2 Small drops:

When droplets grow due to adsorption they are perfectly circular (figure (3,b)), so they can be recognized by Hough transform. In the binary image that is obtained after eliminating the pillars, A_{tot} can be calculated, so in order to get the number of drops, it is just needed to separate the overlapping droplets. For this purpose, the binary image shown in figure (4,a) is turned into a distance map using the Euclidean distance as is shown in figure (4,b). The distance map shows the points inside the droplets with maximum distance from the background area. Therefore, for two overlapping droplets the distance map will eliminate the area of intersection because it is close to background and show two separate points related to droplets centers. In other words, the distance map shows just the droplets centers separately and can be used as an efficient tool for separating overlapping droplets. Now the watershed technique can be applied to binarize the gray scale image. A labeling step enables to get the drop number and finally, equation (1) gives the mean radius of drops.

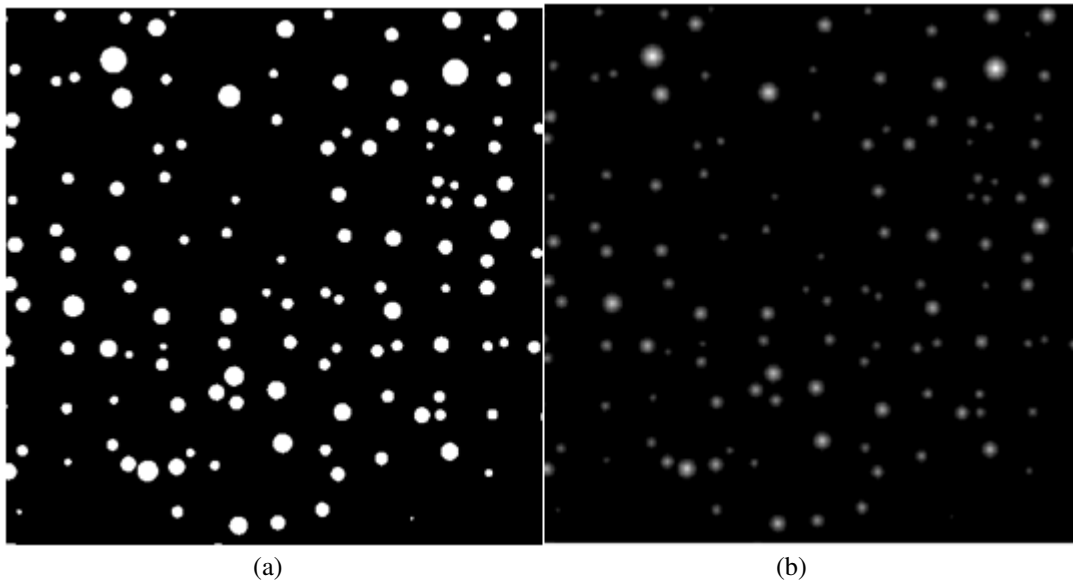


Figure (4): (a) Binary image of small drops, (b) Distance map of small drops. The distance map can separate the overlapping droplets by emphasizing on the points that are in longer distance from contours of droplets.

3.3 Medium drops:

Medium drops form a continuous cluster around the pillars. The problem with these droplets is that their centers are very similar to the background region (Figure (3,c)). So, at first the drops centers should be determined by some criteria. The droplets centers are different from background region from two points of view, they are more convex, but they have more spread gradient magnitude. Therefore, the original images are firstly thresholded by Otsu's method to get brighter regions. Then two techniques are applied to sufficiently large white regions. The regions that correspond to a convexity rate under 4% figure (5,a) or the ones which correspond to a low rate of gradient magnitude according to figure (5,b) are related to drops centers. The limited rate of gradient magnitude corresponds to 4/5 of the mean gradient magnitude of the neighborhoods. Finally, the binary images are made by applying the watershed method.

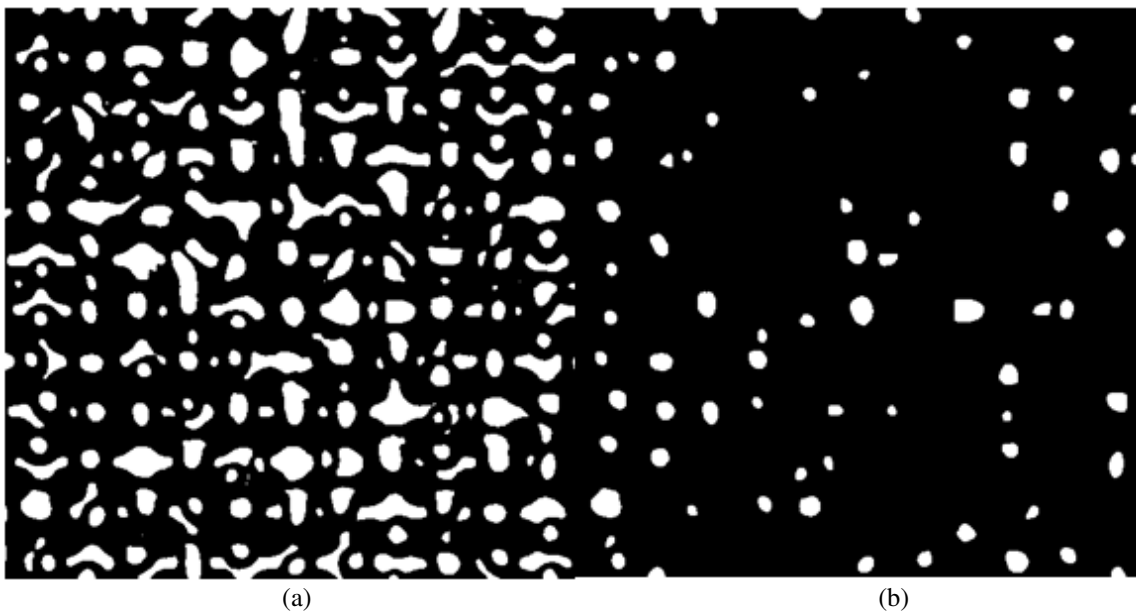


Figure (5): (a) Low gradient part of medium drops, (b) Convex part of medium drops

3.4 Big drops:

Big drops can cover entirely several pillars as shown in Figure (3,d). For binarizing such images, shrinking and gradient properties must be applied. At first, the images are thresholded by Otsu's method, and then the black parts are shrunk in order to get their skeletons around the white regions, which correspond to either drops centers or the background. As was explained for the medium drops, the average of gradient magnitude is calculated on each white region and then. The distribution of this value shows two main peaks related to intensity of background and intensity of droplets centers. The median between these two peaks can be used as an thresholding value for gray scale image. However, when several white regions belong to the same drop, the watershed technique cannot be applied directly. To solve this problem, the white regions are dilated and overlaid with the pillars image. Thus, the white regions which belong to the same drop are connected, then applying a test on convexity enables to reconstruct each drop center.

Finally, a watershed technique is used to detect each drop region that enables to get the drops number and mean radius.

4. ALGORITHM FOR SIMULATING DROPLETS NUCLEATION AND GROWTH

As was explained before, the medium and big droplets are not perfectly spherical and they can form several connected ellipses lying between the pillars. So the algorithm that is proposed here, simulates nucleation and growth of ellipsoidal droplets on the pillared surfaces. At first random ellipses with density of $3.7 \times 10^7 \text{ m}^{-2}$ are distributed on the three dimensional pillared substrate. The height of droplets on the three dimensional substrate is calculated based on their distances with pillars:

- 1) If center to center distance of droplet and pillar $>$ radius of the pillar, meaning that the droplet nucleates on the surface of substrate, so $z = 0$.
- 2) If center to center distance of droplet and pillar = radius of the pillar, meaning that the droplet nucleates on the top of the pillars, so $z = z_{\text{pillar}}$.
- 3) If center to center distance of droplet and pillar $<$ radius of the pillar, meaning that the droplet nucleates on the side of the pillars, so $0 < z < z_{\text{pillar}}$.

Substrate area is considered equal to $1.3 \times 10^{-5} \text{ m}^2$, covered by pillars with radius of $25 \mu\text{m}$, height of $10 \mu\text{m}$ and border to border distance of $12.5 \mu\text{m}$. Each ellipse has three unequal radius considered as: a , $b=e.a$, and $d=f.a$ along the axis X , Y , and Z , where e and f are the random numbers. All the droplets then start to grow thorough an iterating loop. At each iteration, there is an inner loop in which the droplets that are in touch with pillars are recognized, because if a droplet touches a pillar, they will stay in touch until the end of the process and this will be important during coalescence step. After this inner loop, each droplet grows by adsorbing water molecules from humid air, then the droplets that are in touch in each of the planes X - Y , X - Z , or Y - Z unite and form a bigger droplet called daughter drop. The position of daughter drop is identified as follow:

- 1) If both or none of the parents touch at least one pillar, the daughter drop will be in the mass center of its parents.
- 2) If just one of the parents touches a pillar, then the daughter drop will locate at its center point.

The result of coalescence is producing more vacant area on the substrate, on which the small ellipses can nucleate at each iteration. Finally, if a droplet is big enough to slide from the surface, it will leave the surface and clean off other droplets on its path.

This algorithm was written in Matlab environment and took 2 hours for printing the final results.

5. RESULTS AND DISCUSSION

Figures (6) and (7) compare the results of droplets density (N_t) and mean radius (r_{ave}) between real images and droplets generated during the computer simulation. As it can be seen from these images, at first the surface is covered by huge amount of very small droplets. After a while the droplets start to grow by adsorbing water molecules and coalescence, this will result in rapid changes in the graphs of r_{ave} and N_t . This phase relates to small and medium droplets. In this step the shape of droplets changed from circular to more ellipsoidal drops. By increasing the size of droplets, new small droplets start to nucleate in the vacant area produced during coalescence. These small droplets will balance the number and average size of droplets as well as percentage of area occupied by the droplets. Therefore in the final stages, an approximately constant pattern in the graphs of r_{ave} and N_t is visible.

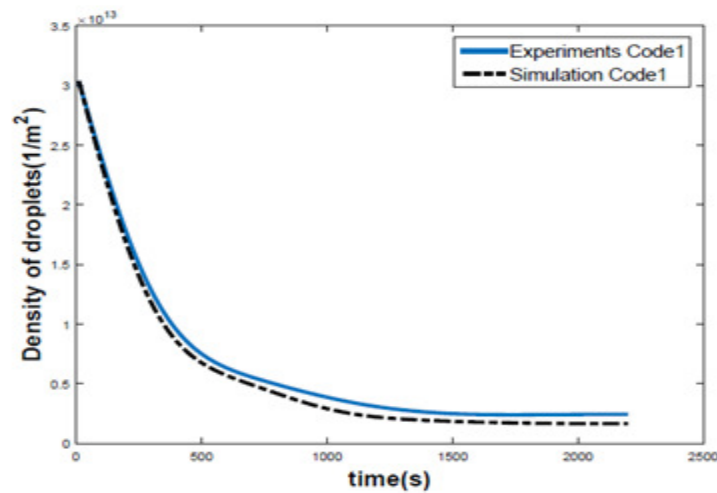


Figure (6): Comparison between density of simulated and real droplets

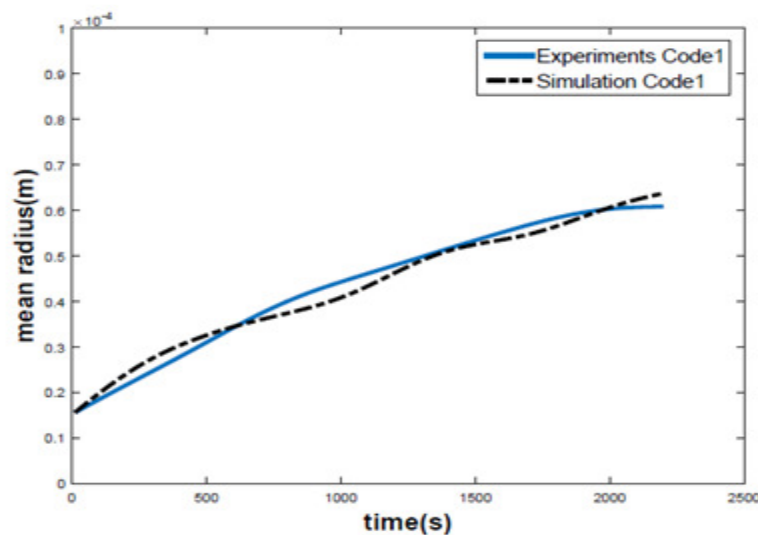


Figure (7): Comparison between mean radius of simulated and real droplets

Figure (8) shows different stages of ellipsoidal droplets growth including initial nucleation, droplet growth due to adsorption, droplet growth due to coalescence and nucleation of new droplets. Figure (8,a) corresponds to nucleation of initial droplets that are introduced as very small droplets, in this earlier stages the percentage of surface covered by liquid is very small and the images are more black. The initial droplets are more circular and their radii a , b , and d are almost the same. Then these droplets start to grow by adsorbing water molecules from humid air and turn into small droplets that are shown in Figure (8,b). Medium and big droplets (Figure(8,c) and (8,d)) are droplets that grow mainly due to coalescence and consequently the area occupied by droplets increases significantly. These droplets that are more ellipsoidal are able to cover the area of several pillars.

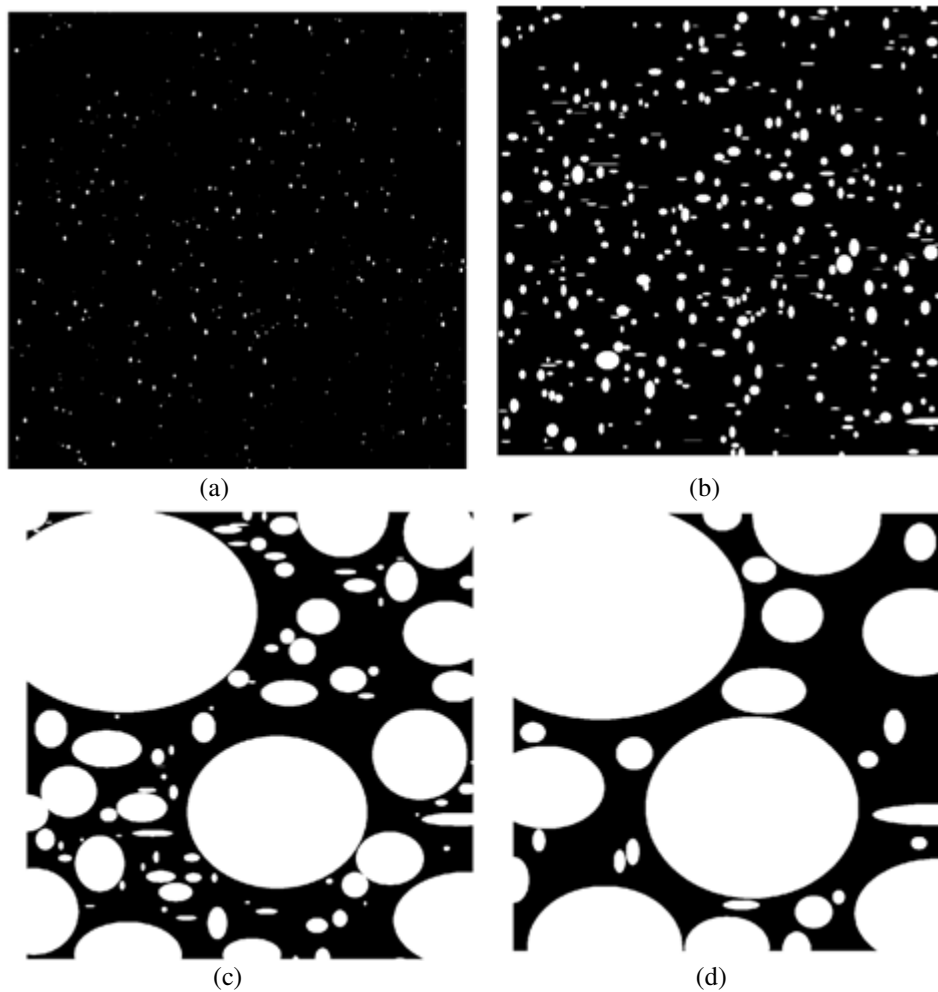


Figure (8): Different stages of ellipsoidal droplets growth. (a) distribution of initial droplets (b) droplets growth due to adsorption, (c) nucleation of new small droplets in the vacant area of the substrate, and (d) droplets growth due to coalescence.

6. CONCLUSION

In this research we developed an image segmentation method for extracting radius and density of droplets growing on a textured surface from gray scale images. The method divides droplets in to four groups based on their radius and apply different segmentation methods for each group. Then the binarized images produced with this algorithm are used to calculate radius and density of droplets. An algorithm for generating ellipsoidal droplets on textured surfaces then is proposed that is able to check coalescence of droplets in three dimensions. The results of this algorithm are in acceptable accordance with the data extracted from binarized images.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of Stephane Valette, Nicolas Pionnier, Remi Berger and Elise Contraires from Ecole Centrale Lyon, who provide the image sets.

REFERENCES

- [1] Baojin Qi, Jinjia Wei, Li Zhang, & Hong Xu, (2015)“A fractal dropwise condensation heat transfer model including the effects of contact angle and drop size distribution”. *International Journal of Heat and Mass Transfer*, Vol. 83, pp259–272.
- [2] Paul Meakin. (1992)“Droplet deposition growth and coalescence”, *Reports on Progress in Physics*, Vol. 55, No. 2, pp157.
- [3] Xue-Hu Ma, Tian-Yi Song, Zhong Lan, & Tao Bai.(2010) ”Transient characteristics of initial droplet size distribution and effect of pressure on evolution of transient condensation on low thermal conductivity surface”, *International Journal of Thermal Sciences*, Vol.49, No. 9, pp1517–1526.
- [4] Maofei Mei , Feng Hu, & Chong Han , Yanhai, (2015) “Time-averaged droplet size distribution in steady-state dropwise condensation”, *Cheng International Journal of Heat and Mass Transfer*, Vol. 88, pp 338–345.
- [5] leon r. glicksman & andrew w. hunt, JR , (1972) ”Numerical simulation of dropwise condensation” ,*International Journal of Heat and Mass Transfer* , Vol. 15, pp. 2251-2269.
- [6] S. Vemuri & K.J. Kim, (2006) “An experimental and theoretical study on the concept of dropwise condensation”, *International Journal of Heat and Mass Transfer*, Vol.49, pp649–657.
- [7] Basant Singh Sikarwar, Sameer Khandekar & K. Muralidhar. (2012),“Mathematical modelling of dropwise condensation on textured surfaces”, *Sadhana*, Vol.38, No.6, pp1135–1171.
- [8] Soo-Chang Pei & Ji-Hwei Horng. (1995), “Circular arc detection based on hough transform”, *Pattern recognition letters*, Vol.16, No.6, pp615–625.
- [9] Xiaoran Yu, Dongchang Xing, Tatsuya Hazuku, Tomoji Takamasa, Takaski Ishimaru, Yuji Tanaka, and Tatsuro Akiba. (2009),”Measurement technique for solid-liquid two-phase flow using normal-line hough transform method”. In *Journal of Physics: Conference Series*, Vol.147, pp 012053. IOP Publishing.
- [10] T. J. Atherton and D. J. Kerbyson. (1993), “Using phase to represent radius in the coherent circle hough transform”, In *Hough Transforms, IEEE Colloquium on*, pp5–1. IET.

- [11] R. Muthukrishnan and M. Radha. (2011), "Edge detection technique for image segmentation", *International journal of recent trends in engineering*, Vol.1, No.2, pp250–254.
- [12] R. Lindken and W. Merzkirch. (2002), "A novel piv technique for measurements in multiphase flows and its application to two-phase bubbly flows", *Experiments in fluids*, Vol.33, No.6, pp814–825.
- [13] S. Beucher and F. Meyer. (1990), "The morphological approach to segmentation; the watershed transformation", *Optical Engineering-New York-Marcel Dekker Incorporated*, Vol.34, pp433–433.
- [14] X. Zabulis, M. Papara, A. Chatziargyriou, and T. D. Karapantsios. (2007), "Detection of densely dispersed spherical bubbles in digital images based in template matching technique - application to wet foams", *Physicochemical and Engineering Aspects*, Vol.309, No.1, pp96–106.

PREDICTING SOFTWARE LAUNCH READINESS IN A COMPLEX PRODUCT

Abhinav Sharma

HCL Technologies, Welwyn Garden City, UK

ABSTRACT

A simple model used successfully for estimating and tracking software defects to predict launch readiness of software in a complex product is described in this paper. The model is based on tracking the number of defects estimated to be found, actually found and resolved to measure the quality of the product. Defect estimates can also help identify quality and process issues in the development and testing phases.

The defect estimation tracking method described here covers the whole project and is split into the three phases Initial Defect Estimates (based on historical data), Interim Revised Estimates (based on actual performance of the project) and Final Defect Tracking (based on testing still to do). The method is based on existing development processes of the team so is easier to implement and has been successfully applied in several projects.

KEYWORDS

Software Reliability Growth Model, Defect Estimation, Software Quality Tracking, Schedule Prediction

1. INTRODUCTION

Almost every project team wants to meet their schedule and cost targets. Delay in a project can waste a lot of resources and may even result in cancellation of the project. In some cases a failed project can also make a company go bankrupt [1]. Early projection of when a project would complete with a quality product also enables the rest of the supply chain to align with the delivery of the product.

This document describes the model used to predict launch readiness of software in a complex product. After providing some background the paper explains the model used to predict launch readiness. This is then explained with metrics from real projects. Finally, issues which need further research and current best practices to adopt are briefly discussed.

To maintain confidentiality, the company, product and project names are not used in the paper and the dates are changed as a further security measure.

2. BACKGROUND

This paper covers complex projects with schedule ranging from six months to about two years. The complexity here is best explained by a) source lines of code (kSLOC) which was over a million, b) interaction of sub-systems which were mechanical, electronics and software and c) the developed software which included embedded, back-office and personal computer applications and low level drivers.

Low confidence in predictability of software launch readiness means that teams in different geographies, including manufacturing, marketing cannot plan to complete brochures and other marketing material to meet the launch date. To improve confidence in predicting launch readiness several software process improvements were initiated and a model to predict launch readiness was developed.

During development, product testing happens at different levels, like unit test, module test, system test, certification test and in different geographic locations. Any of these tests (excluding unit tests) can identify defects which are then included in the model. The reliability of the overall system is tracked using a different mechanism which is not covered in this paper. However, the software issues identified during system testing are treated as defects and are covered in the model.

3. PRIOR PROCESS IMPROVEMENTS

The model described here depends upon the important improvements in defect management which had been implemented within the team in earlier projects. These were:

- Defect Attributes,
- Defect Lifecycle, and
- Defect Tracking

3.1 Defect Attributes

The defect attributes were defined to manage the defects consistently across the teams. The key attributes relevant to this paper are listed below.

- Priority – business priority for fixing the defect as Critical, Major or Minor.
- Severity – severity of the defect from the customer’s point of view as 1, 2, 3 and 4.
- State – defines the lifecycle state the defect is currently in. See section 4.2 for the states defined for the Defect Lifecycle used in the model.

3.2 Defect Lifecycle

The following main states of a defect were defined and then used consistently throughout the projects. The main states relevant to this model are listed below.

- New – the defect is created in this state.
- Assign – the defect is assigned to an engineer to resolve.

- Reject – the defect is rejected as invalid.
- No Action Planned (NAP) – the defect is accepted as valid but will not be fixed.
- Fixed – the defect has been fixed.

3.3 Defect Tracking

Trend charts were used in the earlier projects to track the number of defects created and resolved by week. Although the charts had basic information they provided a simple indicator on whether the rate of finding defects is slowing down or not and if the fix rate is keeping up and closing the gap or not.

4. ISSUES WITH EARLIER PROJECTS

The defect management process worked well but had limited business value in that it did not help in planning for the launch of the products. It was not possible to predict launch readiness of a product to plan related marketing and supply activities. This problem is illustrated by the following two charts.

Figure 3 below shows the Defects Projection and Tracking chart for an older project close to launch. The vertical axis shows the total number of defects found or fixed so far in the project and the horizontal axis is date (which has been changed for reasons of confidentiality). Due to the limited information the development team could not predict when the product would be ready for launch and even at the code freeze date had a significant number of open defects. The project launch in this case was delayed due to high level of open defects.

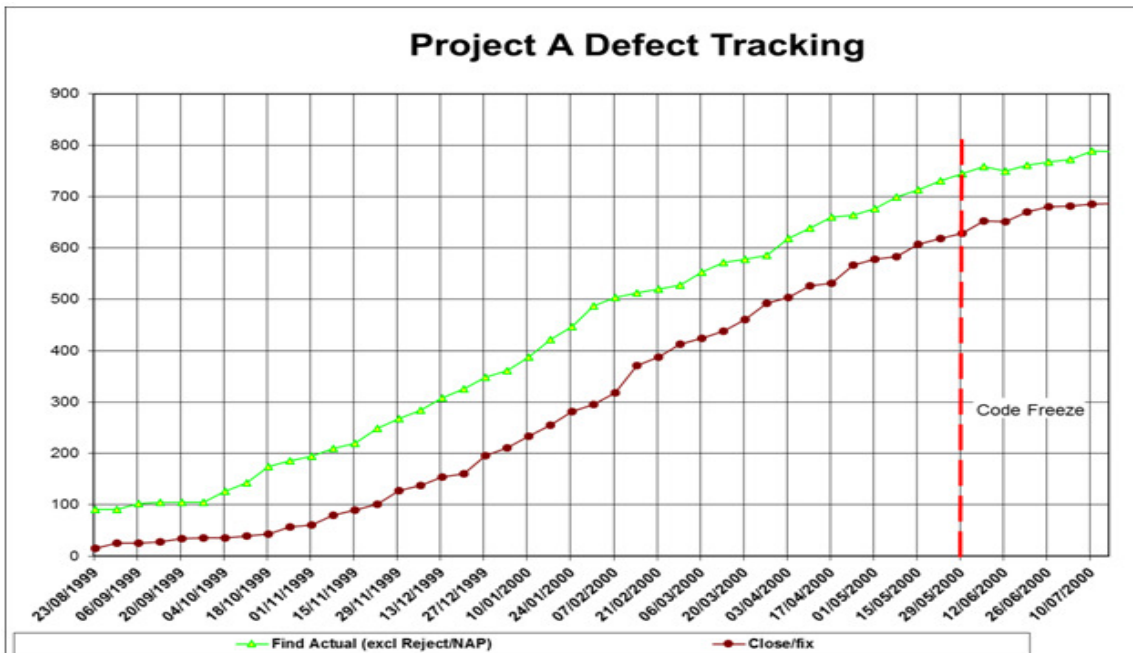


Figure 3 Project A – Defect Projection and Tracking

A few improvements were made to the earlier defect find and fix trend charts. The charts now had a target number of defects to be fixed and showed the maximum fix capability and minimum fix capability of the development team during the hardening period. This helped the defect tracking further by ensuring that the fix rate was kept close to the maximum fix capability.

The improved chart is shown in Figure 4 below. The vertical axis shows the total number of defects found or fixed so far in the project and the horizontal axis is date (which has been changed for reasons of confidentiality). The chart also shows that in the middle of the testing when the fix rate started to slow down corrective actions were taken to bring it closer to the maximum fix projection rate. The chart shows the corrective actions identified while tracking

- 2 spikes show where corrective actions to increase defect find rate were applied and
- one spike shows where corrective action to increase the defect fix rate during the system testing was applied.

Although the updated defect projection and tracking chart was more useful than before but the ability to predict launch readiness was still not there. Launch of this project was also delayed due to the large number of open defects. This eventually led to the development of the software launch readiness prediction model described in this paper.

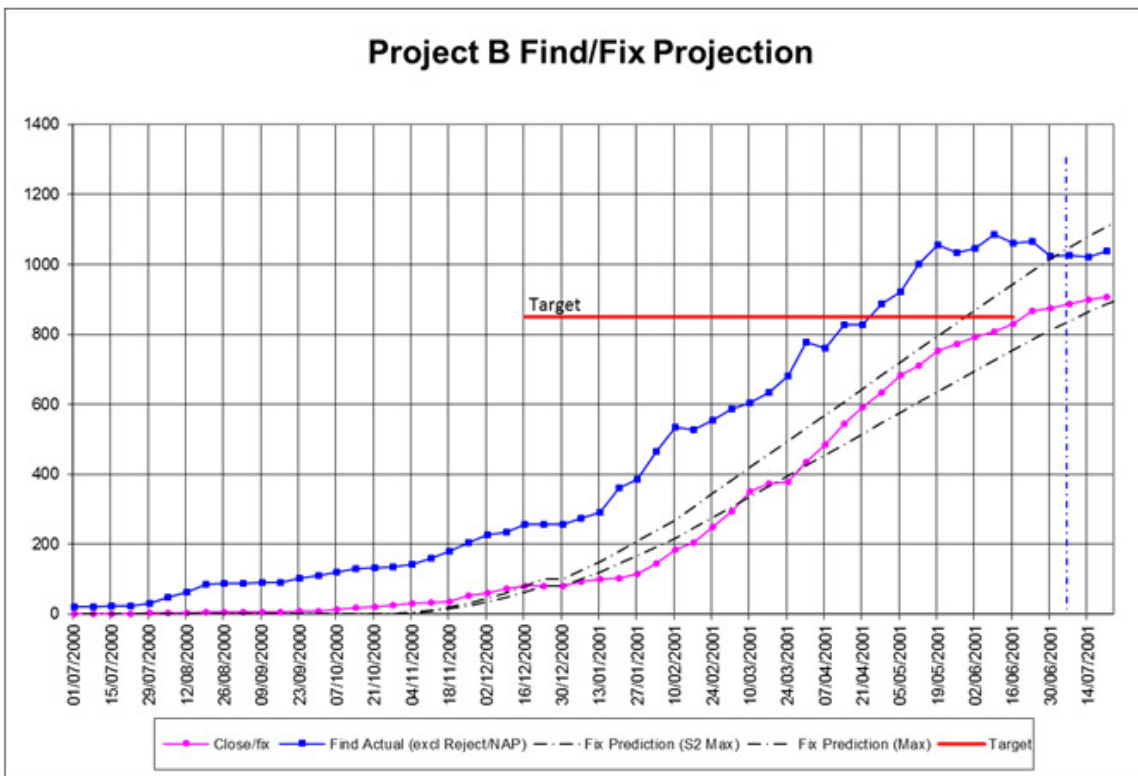


Figure 1 Project B - Defect Projection and Tracking

5. SOFTWARE LAUNCH READINESS PREDICTION MODEL

The following sections describe the key concepts used in the model and the software launch readiness prediction model in detail.

5.1 Concepts Used in the Model

In this model a defect is defined as an error in the software programme, that when executed under particular conditions will result in a failure. Failure means that a function of the software does not meet user requirements.

Reliability is usually defined as the probability that a system will operate without a failure for a specified time under specific operating conditions. Reliability is concerned with the time between failures or its reciprocal, the failure rate. In this model the reliability is tracked as Failure based, where, cumulative failures are recorded within a given time interval (of a week.)

Reliability can also be tracked as system shutdown and system reset rates separately. Any software related shutdowns and system resets are treated as high priority defects. So the assumption here is that if the defects are fixed in time then the rest of the reliability metrics will also get under control.

5.2 Software Development Lifecycle

The Model proposed here uses historical data in the Initial Phase so the software development processes used in the past projects are important. Minor changes in the processes can be ignored but significant process changes mean that the historical data may be of little use for projection.

The software development lifecycle used is also important to understand the scope of the model. For example,

Delivery: Incremental delivery? Iterative delivery? Software Reliability Growth Model (SRGM) are used in the model for projecting the defect find rate so it is important to pick the right model. Several SRGM differentiate between completion of implementation and start of system testing (hardening). Some SRGM do allow for test to start in parallel with implementation but the model used here assumes that over eighty percent of the development has been completed.

Defects: Lifecycle of a defect is important for the tracking part, for example, when are the defects raised and how are these counted? A forum is used to review and accept failures found in any testing as software defects using a consistent method of classifying, prioritising and allocating defects.

Review: Existing code review/inspection processes? The software quality and by inference estimate of defects depends on the quality and consistency with which review processes are used during development.

5.3 Defect Estimating and Tracking Approach

The approach in the model is based on three phases:

- The Initial Phase: Initial Defect Estimates & Fix Projection
- The Interim Phase: Revised Estimates & Update Fix Projection
- The Final Phase: Defect & Fix Projection with Test Tracking

5.3.1 The Initial Phase: Initial Defect Estimates & Fix Projection

Early in the development, estimate the total defects which will have to be fixed to meet the quality target. There are several methods to estimate defects and my recommendation is to first estimate the size and then use the size to predict the number of defects. This is because, given the project requirements, the 'size' of the project should not change. Also, if the size can be measured on delivery then it makes it easier to update the estimates. Unfortunately, there is no perfect metric for measuring size so for simplicity kSLOC (kilo SLOC) is used in the model for size. However, it is possible to estimate defects directly. In case the product size in kSLOC is estimated then using the historical data estimate the number of defects injected for every 1 kSLOC code developed is determined.

Now all this is put together:

- Given the total defects to find use the Rayleigh curve (or another SRGM [2]) to project defect find rate.
- From the available team estimate when engineers will complete implementation and start fixing defects.
- Determine average fix capability from historical data, as number of days to fix a defect per engineer.

This provides a chart showing the projected defect find and projected defect fix curves. Add the defect estimate and the launch build date as targets on the chart. Finally, add actual find and fix charts which will be populated regularly as part of tracking.

5.3.2 The Interim Phase: Revised Estimates & Update Fix Projection

Once majority of the development is complete, say 80% delivered using Earned Value Method (or another appropriate method), measure the actual size (kSLOC) delivered and defects raised to revise the defect estimate. This is the time to revise the estimates based on implementation already completed, testing started and implementation still to complete.

Now update the following:

- The total defect estimate,
- Fix capability of the team from fixes delivered so far (if significant defects have been delivered) and
- Team availability for the rest of the duration of the project.

And obtain the following from the defect tracking database:

- Defects already found and
- Defects already fixed

Update the chart with the revised find and fix projections. This chart allows for tracking the rate with which defects are being found and fixed and to take corrective actions as required to stay on schedule:

- If defects find rate is low then possible options are to review and improve the test plan, increase test resources etc.
- If defect fix rate is low then possible options are to review and improve the defect fix process, increase engineers allocated to fixing defects etc.

5.3.3 The Final Phase: Defect & Fix Projection with Test Tracking

Once significant testing has completed then switch to using the defects arrival rate from different tests and the amount of testing still to complete to estimate the remaining defects to find.

The first step is to review the total defect estimate made in the Initial Phase by comparing the defects already found and fixed with the estimates. From the defects found so far determine the rate with which different test groups were identifying defects. Using this rate of defect detection and remaining tests still to complete, determine the updated estimate of the remaining defects to find. Update the defect projection and tracking chart with the revised estimates providing the remaining defects to find.

6. CASE STUDY

The model has been applied to the projects following Spiral Model and Scrum adapted for the organisation. The model can also be applied to other lifecycles, for example, Waterfall lifecycle provided the historical data used for estimates and forecasts followed similar lifecycles.

6.1 Projects Selected for the Case Study

The projects were selected from two product families. All the selected projects involved changes in electronics, mechanicals and software. Software size for all the products was very similar and grew over time as new projects added more features. Newer features and thus projects tended to be more complex with higher interaction between different components.

The names and dates of the projects have been changed. The main discussion below is for project C. Project D is an earlier project whose data is used as historical data. The final section provides the charts from other projects where the same model was also used.

6.2 Initial Phase

In project C, the current project, the defects were estimated using the analogous estimate method where the defects in similar modules of an earlier software project (Project D) were used as a starting point. The %Injected factor was derived for each module using the cost factors defined in COCOMO [1] as a guide.

Table 1 Project C - Analogous Defect Estimates

| Module | Project D | %Injected | Most Likely | Minimum | Maximum | Weighted Estimate |
|--------|-----------|-----------|-------------|---------|---------|-------------------|
| A | 130 | 1.5 | 195 | 25 | 220 | 171 |
| B | 140 | 1.2 | 168 | 60 | 200 | 155 |
| C | 234 | 0.8 | 187 | 125 | 200 | 179 |
| D | 53 | 0.25 | 13 | 10 | 25 | 15 |
| E | 60 | 0.25 | 15 | 10 | 30 | 17 |
| F | 46 | 6 | 276 | 200 | 350 | 276 |
| G | 16 | 2 | 32 | 25 | 40 | 32 |
| H | 20 | 1 | 20 | 10 | 25 | 19 |
| I | 8 | 3 | 24 | 8 | 30 | 22 |
| J | 123 | 0.05 | 6 | 5 | 10 | 7 |
| Total | 830 | Total | 937 | | Total | 892 |

The table above shows the number of defects found in Project D, estimated percent defects which will be injected in Project C which then provides the most likely estimate for the defects. The minimum, maximum and most likely estimates for the defects were derived in consultations with the respective SMEs. The weighted estimate is given by the following equation.

$$\text{Weighted Estimate} = \frac{\text{Minimum} + 4 * \text{Most Likely} + \text{Maximum}}{6}$$

6.3 Intermediate Phase

Reasonable calibration using data from the project is now possible as the project is about half way through the testing. Use the development teams' capabilities to fix defects so far to project the fix trend for the rest of the project duration. Similarly, use the existing defect find trend to project the defect find trend for the rest of the project duration.

The chart in Figure 1 shows the defect find and fix projections with the actual defects found and fixed in the early stages of system testing. The vertical axis shows the total number of defects found or fixed in the project to date and the horizontal axis is date (which has been changed for reasons of confidentiality).

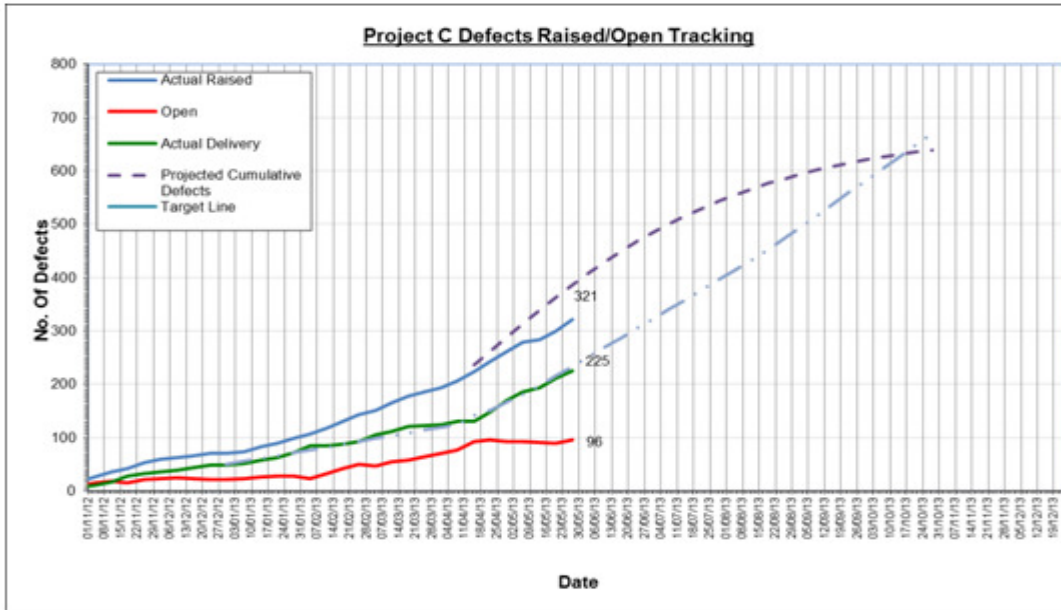


Figure 1 Project C - Defect Projection and Tracking

6.4 The Final Phase

The chart in Figure 2 shows the projected arrival rate of the defects based on the testing still to complete. The vertical axis shows the total number of defects found or fixed to date in the project and the horizontal axis is date (which has been changed for reasons of confidentiality).

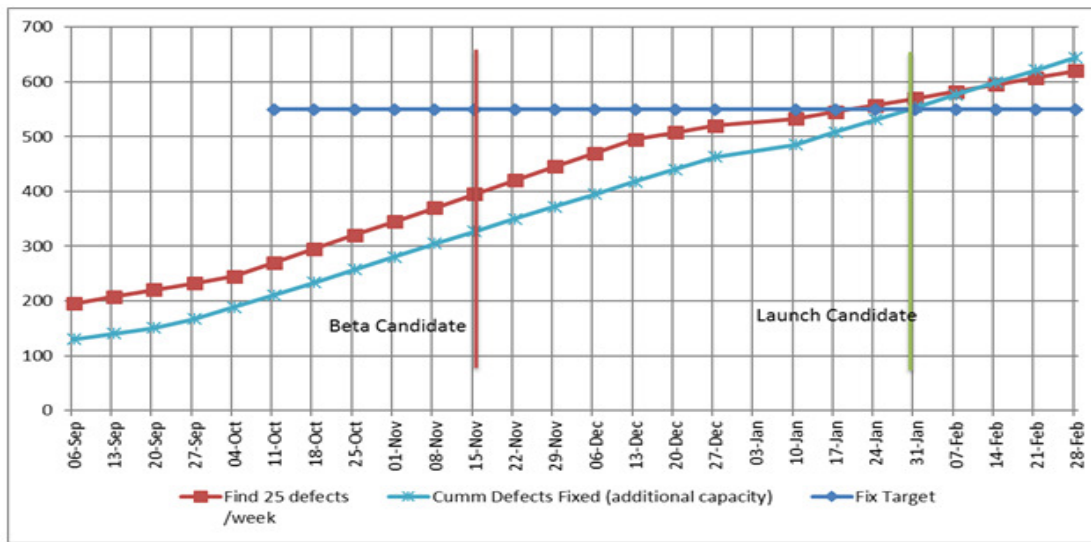


Figure 2 Revised Projection based on Test to Complete

For Project C, discussed earlier, the Table 2 below shows the initial estimates and the defects already found in testing which then provides the number of defects still to find. This can also be used as a sanity check for the estimate of defects still to find based on the defect arrival rate from different tests in progress.

Table 2 Review of Estimates in the Final Phase

| Module | Project D | %Injected | Most Likely | Minimum | Maximum | Weighted Estimate | Project C Found | Still to Find |
|--------|-----------|-----------|-------------|---------|---------|-------------------|-----------------|---------------|
| A | 130 | 1.5 | 195 | 25 | 220 | 171 | 89 | 82 |
| B | 140 | 1.2 | 168 | 60 | 200 | 155 | 125 | 30 |
| C | 234 | 0.8 | 187 | 125 | 200 | 179 | 145 | 34 |
| D | 53 | 0.25 | 13 | 10 | 25 | 15 | 11 | 4 |
| E | 60 | 0.25 | 15 | 10 | 30 | 17 | 12 | 5 |
| F | 46 | 6 | 276 | 200 | 350 | 276 | 190 | 86 |
| G | 16 | 2 | 32 | 25 | 40 | 32 | 18 | 14 |
| H | 20 | 1 | 20 | 10 | 25 | 19 | 16 | 3 |
| I | 8 | 3 | 24 | 8 | 30 | 22 | 17 | 5 |
| J | 123 | 0.05 | 6 | 5 | 10 | 7 | 2 | 5 |
| Total | 830 | Total | 937 | | Total | 892 | 625 | 267 |

7. APPLYING THE MODEL

The method described in this paper of defect find and fix tracking is fairly simple but there are some assumptions which need to be resolved before the model can be successfully applied. The main assumptions which the model relies on are described below.

7.1 Development Process

The method assumes consistency between development processes used in the current and the previous projects for the Initial Estimation Phase. From this method's point of view, this includes having standard review process, definitions of defects, priority, severity etc. Benefits of process improvements take time to percolate through the system so these should only be considered after their first successful implementation.

7.2 Historical Data

The model assumes that historical data from previous projects is present and applies to the current project. In absence of any historical data industry standard metrics will have to be used in the beginning.

7.3 Tool Calibration

The models and techniques mentioned in this document have been developed under specific environment and need to be calibrated for the organisation.

7.4 Initial Estimate of Total Defects

It is assumed that the team can estimate defects reasonably reliably. Developers find it difficult to estimate number of defects but the method presented here requires the initial estimate of total defects which will be found during development and test. One method which has been found to work better is a mix of analogy and work break-down system (WBS). In this method break the system into smaller sub-systems (WBS items) and then compare the new project with the defects fixed in similar projects delivered in the past.

7.5 Consistent Allocation of Defects

An important assumption is that the team is disciplined in using the related processes throughout the software development lifecycle. Relation between failure and defect is often unclear so a forum which is consistent in its analysis of the incoming failures may help the team. This forum also needs to ensure that duplicate defects are not assigned but are rejected instead.

8. FURTHER RESEARCH WORK

Research in SRGM has included adaptations to existing models to take differences in development and testing processes, for example, developing and testing phases in parallel, restarting test after fixing defects, into account. Artificial intelligence is also being used, to improve learning from the historical database which can then be used to predict schedule and quality of future projects.

Constructive Cost Modelling [1] (COCOMO) was developed by Barry Boehm and is used by several researchers and tool vendors, for example, COSTAR [7], Cost Xpert[4]. Software Lifecycle Management by QSM [3] uses historical data, Raleigh distribution to manage complete software planning and tracking tool SLIM-Suite. Bayesian Belief Network is another way to predict software (or product) quality (and risks) which AgenaRisk [6] is using.

Further research is required in the following main areas.

- Reliable early defect estimates with limited information
- Machine learning techniques to improve estimates and predictions
- Data mining of the defects database for estimates and predictions

9. CONCLUSIONS

Delivering reliable software on schedule is a concern in all development organizations. With increase in size and complexity, several vendors and research institutes are looking into tools and methods on how to improve predictability in software development. Most of these tools are expensive and require significant effort to learn and normalize for the development teams to start getting benefits.

The method described in this paper is simple and practical and any team, disciplined to use processes consistently can start using it with little additional effort. However, initially, the team will need some historical data to base their estimates and predictions on. The method uses tools and concepts readily available to all. The method has been successfully applied in several projects in the past with excellent results. The method can be easily enhanced as more and more data is collected within the development teams without tying them to an external vendor.

REFERENCES

- [1] Robert N. Charette, Why Software Fails, 2008, <http://spectrum.ieee.org/computing/software/why-software-fails>, last accessed on 11th August 2017.
- [2] Reliability Growth Model, <http://www.ece.uvic.ca/~itraore/seng426-07/notes/qual07-8.pdf>, last accessed 11th August 2017.

- [3] Barry W Boehm et al, Software Cost Estimation with COCOMO II, 2000, ISBN-10: 0137025769
- [4] CoStar, <http://www.softstarsystems.com/>
- [5] Cost Xpert, <http://www.costxpert.com/>
- [6] Software Lifecycle Management (SLIM), <http://www.qsm.com>
- [7] AgenaRisk, <http://www.agenarisk.com>
- [8] P.K. Kapur, D.N. Goswami, Amit Bardhan, Ompal Singh, Flexible software reliability growth model with testing effort dependent learning process, Applied Mathematical Modelling, Volume 32, Issue 7, July 2008, Pages 1298-1307
- [9] M.Xie, G.Y.Hong, C.Wohlin A Practical Method for the Estimation of Software Reliability Growth in the Early Stage of Testing, <http://www.wohlin.eu/issre97.pdf> last accessed 11th August 2017.

AUTHOR

Abhinav Sharma is a PMP® (PMI certified) professional with more than 25 years of proven track record in Product/Program/Project Management and development. Main products covered Wearable IoT, Consumer products, Multi-function printers and embedded software. Proven track record of delivery to schedule of large, complex, multi-site, multi-discipline projects and implementing lean/agile/process improvement initiatives.



LIGHTWEIGHT KEY MANAGEMENT SCHEME FOR HIERARCHICAL WIRELESS SENSOR NETWORKS

Mohammed A. Al-taha¹ and Ra'ad A. Muhajjar²

¹Department of Computer Science, College of Science, Basrah University, Iraq

²Department of Computer Science, College of Computer Science and
Information Technology, Basrah University, Iraq

ABSTRACT

Wireless Sensor Networks (WSNs) are critical component in many applications that used for data collection. Since sensors have limited resource, security issues have become a critical challenge in Wireless Sensor Networks. To achieve security of communicated data in the network and to extend the WSNs lifetime; this paper proposes a new scheme called Lightweight Key Management Scheme (LKMS). LKMS used Symmetric Key Cryptography that depends only on a Hash function and XOR operation. Symmetric Key Cryptography is less computation than Asymmetric Key Cryptography. Simulation results show that the proposed scheme provides security, save the energy of sensors with low computation overhead and storage.

KEYWORDS

Wireless sensor Networks, Key Management, Symmetric Cryptography, hash function, XOR

1. INTRODUCTION

Wireless sensor networks (WSNs) consist of hundreds, even thousands of low-cost devices called sensor nodes. These sensors have resource constraints such as limited power resource and low memory. Sensors in WSNs can communicate with each other by radio channel to transmit the data to the centric node known as the sink node (Base Station). WSNs has formed the basics for covering a different range of applications such as health care, military, environmental monitoring and other fields [1].

WSNs can be classified as flat networks and hierarchical networks. In flat networks, every sensor in the network has the same characteristics (battery lifetime, storage capacity, processor and transmitted power) and perform the same task. In flat networks sensor nodes can transmit data to its neighbour one by one to the sink. In hierarchical networks, the network divided in to several groups called clusters (each cluster has a head called cluster head and the other nodes called cluster members). Hierarchical WSNs can be implemented as homogeneous and heterogeneous. All sensor nodes have the same capabilities in homogeneous WSNs. In heterogeneous WSNs incorporate different types of sensor nodes that have different capabilities (small number of sensors with powerful characteristics elects as Cluster head and a large number of low characteristics sensors elects as Cluster members).

Most applications that use WSNs [2] have sensitive data as in military applications. Due to limitation resource of sensor node and the hostile environment make it a big challenge to secure the network. Key management is the first requirements to secure communication in WSNs. Key management includes generation, distribution and installation the keys inside sensors and it should support the node addition and deletion with revocation and update the keys [3]. Symmetric and Asymmetric cryptography are being used to achieve security (authentication, integrity, privacy) in WSNs. In Symmetric, each node (sender and receiver) shared the same secret key that used for encryption and decryption the data communicated in the network. In Asymmetric key, each node, the sender and receiver, have two keys which are the public key that known to all nodes in the network and secret key which is private [5]. Symmetric cryptography is less computation and consuming less energy than Asymmetric Key [4].

In this paper, Lightweight Key Management Scheme(LKMS) for heterogeneous WSNs has been proposed to generate symmetric key that use only a hash function and XOR operation to provide security in WSN.

2. RELATED WORKS

Many key management schemes have been proposed to protect the communication among sensors in the WSNs. In [6], the authors proposed symmetric key by using LU decomposition matrix of length $n \times n$ where n represent the total number of sensor nodes. Each node is pre-loaded randomly with one row of L matrix with corresponding column of U matrix, when two nodes want to communicate first they send to each other their L row then each node calculate session key by multiplying the receiving L with its U column and found that they shared the same value, the proposed scheme used Rivets Cipher (RC5) algorithm to encrypt / decrypt of transmitted data. The authors in [7], used Hybrid key management for heterogeneous WSN, they used Symmetric and Asymmetric cryptographic technique in three levels, first level used signature encryption algorithm based on Elliptic Curve Cryptography (ECC) to secure the communication between the sink and the cluster head. In second level, Diffie-Hellman key exchange algorithm based on ECC is used to generate shared key between the cluster head and sensor nodes. finally, Symmetric session key is used between two sensor nodes because the limited resource of the sensor nodes. Each cluster head selects random value r and broadcast it to all cluster members.

The scheme used shared value based on r when two nodes want to communicated. The authors of [8] proposed a key management scheme for hierarchical WSN by using Power aware routing protocol and track – sector clustering, the track sector clustering scheme is used for minimizing the transmission data redundant by means of reducing the connection between the sink node and the cluster head. Hybrid Elliptic Curve Cryptography (HECC) technique is implementing that used 80 bits key size for securing the routing. In [9] the authors proposed a new scheme used three types of keys: Network key, group key and pairwise key. The network key is used for encrypting the broadcast message and for authenticate the new node, group key is used as a shared key between all the sensor nodes in the same cluster, and the last key, the pairwise key is shared between specific pair of nodes. In the proposed scheme, they used assistant node to improve the security and reduce resource cost when cluster head is compromised. In [10], The authors proposed key management for hierarchical WSN using UAV to establish session key between two nodes. The proposed method used symmetric cryptographic key management which is depended on XOR operation and hash function. UAV used as centre unit to reduce the storage. Two nodes can't establish session key only with the help of UAV.

3. THE PROPOSED METHOD

3.1 Network Model

The proposed scheme adopts the Heterogeneous hierarchical WSN as shown in the Figure 1.

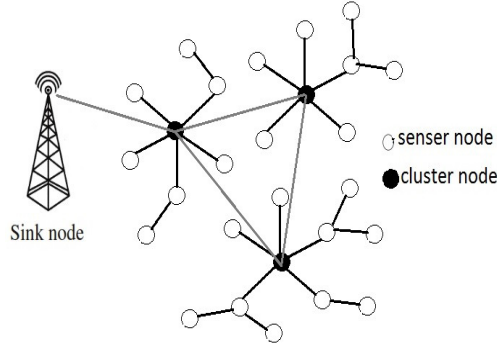


Figure1. Network Model

The proposed scheme assumes that network has the following properties:

- 1- The sink node has unlimited storage and sufficient energy with trusted place.
- 2- All sensor nodes are static.
- 3- If sensor node is captured by attacker, sensors can extract all information from it.
- 4- Cluster head is equipped with tamper resistant. If the cluster head captured by attacker, sensors cannot extract the information stored on it.

The notation of our scheme is summarized in table 1:

Table 1. Notation

| Symbol | definition |
|--------------|---|
| K_{IN} | Initial key from sink node |
| K_{CN} | Shared value between cluster and its member |
| K_{Hi} | Shared value between cluster i and sink node |
| SK_{CH-BS} | Session key between sink node and cluster head i |
| SK_{CH-ab} | Session key between Cluster head-a and cluster head-b |
| SK_{AB} | Session key between Cluster head and sensor node or between sensor node-a and sensor node-b |
| ID_i | Identity of sensor node i |
| ID_{Hi} | Identity of cluster i |
| r | Random value |
| T | Time stamp |
| $hash()$ | Hash function |
| \oplus | XOR |
| \parallel | concatenation |

3.2 Pre-distribution Phase

Before deployment of the sensor nodes, sink node selected initial key K_{IN} . Each node is assigned a unique ID_i and pre-loaded with the initial K_{IN}

3.3 Key Management Phase

After the deployment, each cluster head broadcasts HELLO message which include its ID. Each sensor node may receive messages coming from more than one cluster head. Sensor node chose a cluster head that have best signal strength. Sensors are storing one more parameter which is the cluster head identity. Now each sensor node has three parameters: initial key, its identity and cluster head identity. After that each cluster head selects random value r and broadcasts it to all cluster member and to sink node encrypted by initial key K_{IN} . Each sensor node gets r by decrypting it with the initial key K_{IN} , then each sensor node calculates shared value K_{CN} with its cluster head and all cluster member as follow:

$$K_{CN} = hash(r||ID_H||K_{IN})$$

After that the node will delete r . When sink node get the encrypted r , sink node decrypt it by using K_{IN} and calculating a shared value between the sink and the cluster head CH_i as

$$K_{Hi} = hash(r||ID_{Hi})$$

The proposed scheme has three types of keys; between sink node and CH_i , between cluster head and cluster head and between cluster head and sensor member or sensor node and sensor node in same cluster

3.3.1 Sink node & cluster head

When the cluster head wants to be communicated with the Sink node, cluster head and sink use K_{Hi} as shared value between them. The session key process can be described as below:

- 1- First the Cluster Head send its ID to Sink node
- 2- After that, sink node select nonce random value R_{BS} and compute:

$$X_{BS} = R_{BS} \oplus K_{Hi}$$

$$Y_{BS} = hash(R_{BS}||K_{Hi}||T_1)$$

Then sink node send [X_{BS}, Y_{BS}, T_1] to Cluster Head

- 3- Cluster Head receive [X_{BS}, Y_{BS}, T_1] from sink node, first it verified the time stamp whether $|T_1 - T_C| < \Delta T$ or not. If verification holds then computes

$$R'_{BS} = X_{BS} \oplus K_{Hi}$$

$$Y'_{BS} = hash(R'_{BS}||K_{Hi}||T_1)$$

If $Y'_{BS} = Y_{BS}$ then Cluster Head select a random nonce R_{ch} , otherwise send a rejection message to sink node.

Now Cluster Head computes:

$$X_{ch} = R_{ch} \oplus K_{Hi}$$

$$Y_{ch} = hash(R_{ch}||K_{Hi}||T_2)$$

Then Cluster Head send [X_{ch}, Y_{ch}, T_2] to sink node

- 4- Sink node receive [X_{ch}, Y_{ch}, T_2] from Cluster Head, first it verified the time stamp whether $|T_2 - T_C| < \Delta T$ or not. If verification holds then computes

$$\begin{aligned} R'_{ch} &= X_{ch} \oplus K_{Hi} \\ Y'_{ch} &= \text{hash}(R'_{ch} || K_{Hi} || T_2) \end{aligned}$$

If $Y'_{ch} = Y_{ch}$ proceed further, otherwise send a rejection message to sink node

- 5- Finally, both sink node and Cluster Head agree on same session key

$$SK_{CH-BS} = \text{hash}(R_{ch} \oplus R_{BS})$$

3.3.2 Cluster Head & Cluster Head

When the cluster head want to be communicated with other Cluster Head, they use K_{IN} as shared value between them. The session key process can describe as below:

- 1- First the Cluster Head-a send its ID to second Cluster Head-b
2- After that, Cluster Head-b select nonce random value R_{CHb} and compute:

$$X_{CHb} = R_{CHb} \oplus K_{IN}$$

$$Y_{CHb} = \text{hash}(R_{CHb} || K_{IN} || T_1)$$

Then Cluster Head-b send [X_{CHb}, Y_{CHb}, T_1] to Cluster Head-a

- 3- Cluster Head-a receive [X_{CHb}, Y_{CHb}, T_1] from Cluster Head-b, first it verified the time stamp whether $|T_1 - T_C| < \Delta T$ or not. If verification holds then computes

$$\begin{aligned} R'_{CHb} &= X_{CHb} \oplus K_{IN} \\ Y'_{CHb} &= \text{hash}(R'_{CHb} || K_{IN} || T_1) \end{aligned}$$

If $Y'_{CHb} = Y_{CHb}$ then Cluster Head-a select a random nonce R_{CHa} , otherwise Cluster Head-a send a rejection message to Cluster Head-b.

Now Cluster Head-a computes:

$$X_{CHa} = R_{CHa} \oplus K_{IN}$$

$$Y_{CHa} = \text{hash}(R_{CHa} || K_{IN} || T_2)$$

Then Cluster Head-a send [X_{CHa}, Y_{CHa}, T_2] to Cluster Head-b.

- 4- Cluster Head-b receive [X_{cha}, Y_{cha}, T_2] from Cluster Head-a, first it verified the time stamp whether $|T_2 - T_C| < \Delta T$ or not. If verification holds then computes

$$\begin{aligned} R'_{CHa} &= X_{CHa} \oplus K_{IN} \\ Y'_{CHa} &= \text{hash}(R'_{CHa} || K_{IN} || T_2) \end{aligned}$$

If $Y'_{CHa} = Y_{CHa}$ proceed further, otherwise send a rejection message to Cluster Head -b.

- 5- Finally, both Cluster Head-a and Cluster Head-b agree on same session key

$$SK_{CH-ab} = hash (R_{CHa} \oplus R_{CHb})$$

3.3.3 Cluster Head & Sensor OR Sensor & Sensor

Suppose that two sensor nodes A and B are neighbors. Session key process is presented as bellow:

- 1- First node A send its ID to node B
- 2- Node B select nonce random value R_B and compute:

$$X_B = R_B \oplus K_{CN}$$

$$Y_B = hash (R_B || K_{CN} || T_1)$$

Then node B send [X_B, Y_B, T_1] to node

- 3- Node A receive [X_B, Y_B, T_1] from node B, first it verified the time stamp whether $|T_1 - T_C| < \Delta T$ or not. If verification holds then computes

$$R'_B = X_B \oplus K_{CN}$$

$$Y'_B = hash (R'_B || K_{CN} || T_1)$$

If $Y'_B = Y_B$ then node A select a random nonce R_A , otherwise send a rejection message to node B.

Now node A computes:

$$X_A = R_A \oplus K_{CN}$$

$$Y_A = hash (R_A || K_{CN} || T_2)$$

Then node A send [X_A, Y_A, T_2] to node B

- 4- Node B receive [X_A, Y_A, T_2] from node A, first it verified the time stamp whether $|T_2 - T_C| < \Delta T$ or not. If verification holds then computes

$$R'_A = X_A \oplus K_{CN}$$

$$Y'_A = hash (R'_A || K_{CN} || T_2)$$

If $Y'_A = Y_A$ proceed further, otherwise send a rejection message to node A.

- 5- Finally, both node A and B agree on same session key

$$SK_{AB} = hash (R_A \oplus R_B)$$

4. SECURITY ANALYSIS

The security analysis of the proposed scheme can be discussed as following:

4.1 Key updating

In the proposed scheme, each session used key different from the others. Every session key depends on shared value between two nodes. For more security, the shared value between two nodes should be updated periodically. After certain period, the update phase is started. Cluster

head selects new random value r' that is different from old r and it's never used before, then broadcasts it to all cluster member and to the sink node encrypted with initial key K_{IN} . When the nodes receive the new encrypted r' they get it by decrypting it with initial key K_{IN} . After that, nodes calculate the new shared value K_{CN}' as following:

$$K_{CN}' = \text{hash}(r' || ID_H || K_{IN})$$

Then nodes will delete r' . The sink node gets the new r' by decrypting it with initial key K_{IN} . After that calculate a new shared value between the sink node and the cluster head CH_i as

$$K_{Hi}' = \text{hash}(r' || K_{Hi})$$

4.2 Add New Nodes

The sensor nodes have limited energy and after some time the node will die, for that, the died node should be replaced with new node. Before deployment the new node, the sink node is known which cluster head is belong to according to location of the new node. So sink node pre-loaded the new node with cluster head ID, initial key K_{IN} and the shared value K_{CN} according to the value of r of that cluster.

4.3 Key Revocation

When a node captured, all the security information stored on it will become compromised. The shared value between the sensors in the same cluster should be updated. The cluster selects new random value r' and continue with update phase.

5. PERFORMANCE EVALUATION

We evaluate the performance of our scheme from storage overhead, energy consumption and time. In our simulation environment, we used 100 sensors that randomly deploy with area 100m*100m, 9 sensors as cluster head and 91 as sensor nodes. The transmission range of cluster head is 40m and the sensor nodes is 20m with initial energy 5J and 15J respectively.

5.1 Storage Overhead

In [10] before the deployment, each node is preloaded with its ID and a secret key. After deployment, each sensor node need to store one more shared value α . The total number of this value is fixed by its neighbors. In our scheme before deployment each node is preloaded with its ID and initial key, after deployment each node in the same cluster store one more value K_{CN} . Each sensor node is store just two values in its memory.

5.2 Energy Consumption

The energy consumption of the proposed scheme for transmit 100 packets is less than in [10] as showing in figure 2.

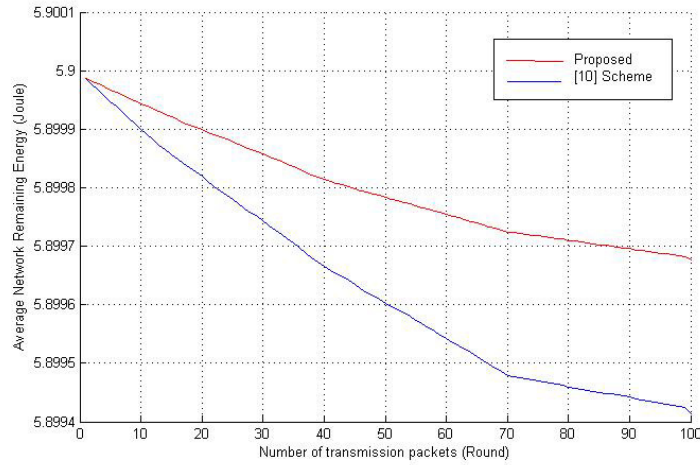


Figure 2. Energy Consumption for transmit 100 packets

Our scheme consumes less energy comparing with [10].

5.3 Time Consumption

The proposed scheme tacks less time to generate session keys and transmit the sensed date. Figure 3 shows the time for transmit 100 packets and compared with [10].

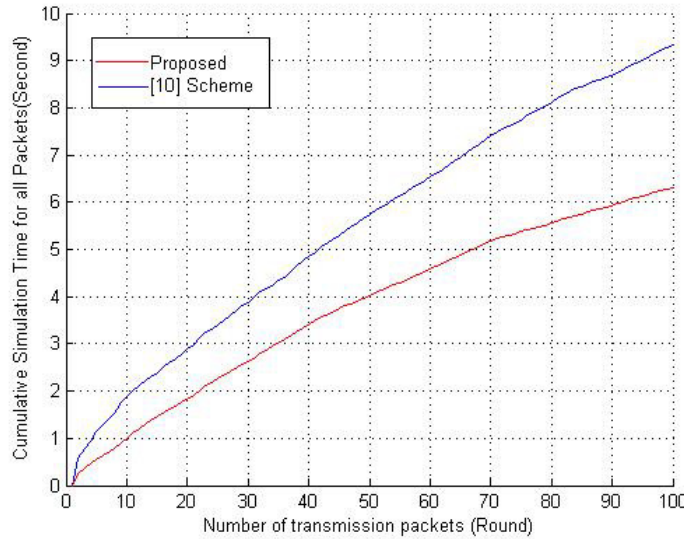


Figure 3. Time Consumption for transmit 100 packets

6. CONCLUSION

In this paper, a Lightweight Key Management Scheme (LKMS) for heterogeneous Wireless Sensor Networks was presented, that used symmetric cryptography only a hash function and XOR operation to establish a session key between any two nodes. Through performance evaluation, we reduce the storage overhead and extend the network lifetime by consuming less energy and take less time to establish secure communication among the nodes.

LKMS use shared value between nodes to establish symmetric session keys and update that value at regular interval to avoid node capture attack and to assure that only legal nodes can be communicated.

Simulation and analysis shows that LKMS has good energy efficient, less time consuming and low storage overhead than other similar schemes.

REFERENCES

- [1] Ahmed A. Alkadhawee and Songfeng Lu, "Prolonging the Network Lifetime Based on LPA-Star Algorithm and Fuzzy Logic in Wireless Sensor Network," World Congress on Intelligent Control and Automation (WCICA), IEEE, 2016
- [2] Pawgasame, W., "A survey in adaptive hybrid wireless Sensor Network for military operations". IEEE, in Second Asian Conference Defence Technology (ACDT), pp. 78-83, 2016
- [3] J. Zhang and V. Varadharajan, "Wireless sensor network key management survey and taxonomy," Journal of Network and Computer Applications, vol. 33, pp. 63-75, 2010.
- [4] Omer K. Jasim, Safia Abbas and El-Sayed M. Horbaty, "Evolution of an Emerging Symmetric Quantum Cryptographic Algorithm", Journal of Information Security, Vol. 6, pp. 82-92, 2015
- [5] Ayman T., Ayman K. and Ali C., "Authentication Schemes for Wireless Sensor Networks," in Mediterranean Electrotechnical Conference (MELECON), IEEE, pp. 367-372, 2014
- [6] Khurajam S. K and Radhika K R " A Novel Symmetric Key Encryption Algorithm Based on RC5 in Wireless Sensor Network" International Journal of Emerging Technology and Advanced Engineering , IJETAE, Volume 3, Issue 6, 2013.
- [7] Ying, Z. and Pengfei, J. "An Efficient and Hybrid Key Management for Heterogeneous Wireless Sensor Networks" ,Control And Decision Conference (2014 CCDC),The 26th Chinese , ,IEEE, 2014
- [8] V Vijayalakshmi, R Sharmila, and R Shalini. "Hierarchical key management scheme using hyper elliptic curve cryptography in wireless sensor networks". In Signal Processing, Communication and Networking (ICSCN), 2015 3rd International Conference on. IEEE, 2015
- [9] Zhang, X., and Wsn, A. C. "An Efficient Key Management Scheme in Hierarchical Wireless Sensor Networks". 2015 International Conference on Computing, Communication And Security (ICCCS), IEEE,doi:10.1109/CCCS.2015.7374122, 2015
- [10] Akansha Singh, Amit K. Awasthi and Karan Singh," Lightweight Multilevel Key Management Scheme for Large Scale Wireless Sensor Network",IEEE 2016

INTENTIONAL BLANK

BLOCK CHAIN BASED DATA LOGGING AND INTEGRITY MANAGEMENT SYSTEM FOR CLOUD FORENSICS

Jun Hak Park, Jun Young Park, Eui Nam Huh

Department of Computer Science and Engineering,
Kyung Hee University, Yongin-si, South Korea

ABSTRACT

Along with the increasing use of cloud services, security threats are also increasing and attack methods are becoming more diverse. However, there are still few measures and policies to deal with security incidents in the cloud environment. Although many solutions have been proposed through research on digital forensics for responding to security incidents, but it is still difficult to prove the integrity of evidence collection and storage in the cloud environment. To solve these problems, in this paper, we propose a blockchain based data logging and integrity management system for cloud forensics. In addition, compare the performance of the proposed system with the other blockchain based cryptocurrency.

KEYWORDS

Cloud Computing, Cloud Forensics, Block chain, Data Integrity

1. INTRODUCTION

Cloud computing is a technology that provides physical resources to users through virtualization technology. Profit of enabling network access to a scalable and elastic pool of shareable physical or virtual resources with self-service provisioning and administration on-demand profit, cloud computing market is getting bigger. Because of these characteristics, the number of users using cloud computing is also increased. However, with the growing cloud computing market, security threats began to grow. Many security solutions for the cloud environment are being researched, it is difficult to apply the existing digital forensic methods because of virtualization technology [1]. When the cloud environment is classified according to the service model, access to some system layers is limited in Software-as-a-Service(SaaS) and Platform-as-a-Service(PaaS) environments, access to that layer is controlled by Cloud Service Provider(CSP). So the log data generated in the inaccessible layer needs to be provided to the CSP through agreements[2]. In traditional digital forensics, investigators have full control over the evidence. However, in a cloud environment, the data centers are distributed globally, Cloud Service Customers(CSC) share physical resources, volatile data that disappears when CSC shut down the instance, virtual network, load balancing and auto scaling for providing seamless service environment. Therefore, it is necessary not only to record data for cloud forensics before a security incident for investigation but also to ensure the

integrity of the log data because it is difficult for the investigator to collect the data directly and collect the data from the remote site.

There are several methods for ensuring the integrity of data, one of which is a blockchain. A technique called blockchain or distributed ledger is being studied as a method for ensuring integrity since the previous block affects the value of the next block. Since all blocks are connected like chain, it is possible to verify the integrity of all past blocks simply by verifying the hash value of the immediately preceding block. In this paper, we describe the need for data logging system for cloud forensics and propose a blockchain based data logging and management system for cloud forensics. The paper is structured as follows. Section 2 review about cloud forensics, blockchain and related works for ensuring data integrity. Section 3 describes the proposed system. Section 4 compare the performance of the proposed system with the some blockchain based cryptocurrency. Finally in Section 5 describes conclusions and suggest future research directions.

2. RELATED WORKS

2.1. Cloud Forensics

Cloud forensic is a branch of forensic science encompassing the recovery and investigation of material found in cloud environment, often in relation to computer crime[3]. According to NIST[4], computer forensics consists of four steps: Collection involves the process of physical acquisition of data. Examination is the process of combing through the data for items of interest. Analysis is the application of the interesting items to the investigative question. Reporting describes the output of analysis. The difference between cloud forensics and traditional digital forensics is the collection and identification steps. Because outsourcing resource is one of characteristics in the cloud computing. For the more improve forensic investigation procedure, such as the storage and transportation of data stored in the cloud server is added. Because this is need to guarantee the reliability of such data confidentiality and integrity of data forensic investigation. The problem of applying the collection and identification method of digital forensics to the cloud environment is that the cloud environment is an outsourcing resource to use the desired service or resource from the CSP and it is difficult to know the actual location of the data because of the virtualization technology applied. Therefore, there is a need for a reliable identification and collection method of data that takes into account the characteristics of the environment.

2.2 Cloud Forensic Challenges

Unlike legacy systems that own all of the computing resources, in the cloud computing environment, the CSP provides infrastructure, platform, application. The CSC utilizes the services provided. This structural difference causes many issues in cloud forensic, such as the storage of data and storage locations, and access the data.

The first reason why difficult to apply forensic technology in cloud computing is that data processing is dispersed in large scale of computing resources. Second, in traditional computer forensics, investigators have full control over the evidence. However, it is very hard in cloud environment. Third, there is a lack of reliable evidence as it is difficult to collect evidence due to the multi-tenant features. Fourth, when VM shut down, it is difficult to preserve volatile data.

Fifth, chains of custody might clearly depict how the evidence was collected, analyzed, and preserved. Sixth, investigators are completely dependent on CSPs for acquiring evidence[5]. In order to solve the problem, it might be saved for prevent loss of volatile data (e.g. snapshot). Moreover, collected evidence might be ensured that the integrity has not been manipulated during the process.

2.3 Previous Research on Data Management Methods in a Cloud Computing Environment

In the cloud environment, data centers are scattered around each country, so there is a possibility that users feel that one data but it is distributed among several physical machines actually. Chun, Byung-Gon et al.[6], propose a method to manage data through a replica set that replicates all data in order to prevent data loss in a distributed node environment and to minimize damage. Although this method has the advantage of solving the data loss problem, there is a disadvantage that the data is managed through the replica, which causes a large maintenance cost. Moreover, since the cloud environment has a service model that pay-as-you-go, it is difficult to apply the method as it is.

Nepal, Surya et al. propose a service that guarantees the integrity of data in cloud storage service[7]. The system provides a way to prevent data tampering in a cloud environment by adding an integrity service provider in a scenario where a cloud service user uses a storage service to upload / download data to / from the cloud, called Data integrity as a Service(DIaaS). This service consists of a Key Management Service (KMS) that manages key values, a Trust Management Service (TMS) that ensures trust, and an Integrity Management Service (IMS) that manages integrity of data. In addition, they propose a model that can guarantee the integrity of data by categorizing the CSP and IMS into four cases, which are trust and untrust respectively.

Edorado Gaetani et al.[8] propose a block chain based data management method for cloud federation environment based on the European SUNFISH project, to solve security problems such as data management method and data integrity in the cloud federation environment. Intrinsic goal of cloud federation is sharing services among members by creating regulated, secured inter-cloud interactions. In order to define possible threats in the cloud federation environment and solve the problem of performance degradation due to the application of block chain technology, they devised a two-layer blockchain based database structure. First layer ensures adequate performance by lightweight distributed consensus protocol, second layer ensures strong integrity guarantees by PoW based blockchan methods.

2.4. Block Chains

A block chain is a distributed ledger technique in which a plurality of peers manage and store data by mutually agreed rules. The nodes (peers) that want to manage the data participate in the P2P network and each node can verify the integrity of the block. Each peer can create a block, where the block of the first successful peer propagates to all peers, and if all the peers agree that the block is justified, the block is added to all peers. If the new block is properly created, it means that the verification of the previous block is also completed. Therefore, the longer the block length, the higher the reliability of the entire block. Verification of the integrity of a block can also verify that all past blocks are correct by comparing the hash value. However, this does not

guarantee that the block is completely trustworthy, and that it has been acknowledged that it has done a lot of work proofing. Therefore, the more peers participating, the safer it is.

New blocks are created using the Proof of Work (PoW) or Proof of Stake (PoS) method. The PoW method is a task to find a hash value that satisfies a certain condition, and it is operated by adjusting the degree of difficulty for an average of 10 minutes in case of Bitcoin[9]. The PoS method is a method for saving the cost and maintenance cost of hardware equipment and is a concept to solve the problem of PoW method in the field of cryptocurrency[10]. Recently, cryptocurrency has been developed that combines both methods properly due to system maintenance cost and security problems. In addition, research is underway to apply not only cryptocurrency but also the fields that need to guarantee the integrity of data. For example, the blockchain based digital content distribution system[11], using blockchain for medical data access management[12], a framework for preventing double-financing[13], blockchains and smart contracts for the Internet of Things[14] are researched.

3. LOGGING SYSTEM FOR CLOUD FORENSICS BASED ON BLOCKCHAIN CHAIN

As mentioned above, the most important consideration for cloud forensics is “how to collect the data?” In cloud computing environment, CSP need to collect and store their own data, in which case there is a possibility of data manipulation and loss, so that the integrity of the data needs to be guaranteed. Therefore, in this section, we propose a system structure that can guarantee the integrity by blockchain technology while CSP collect data itself.

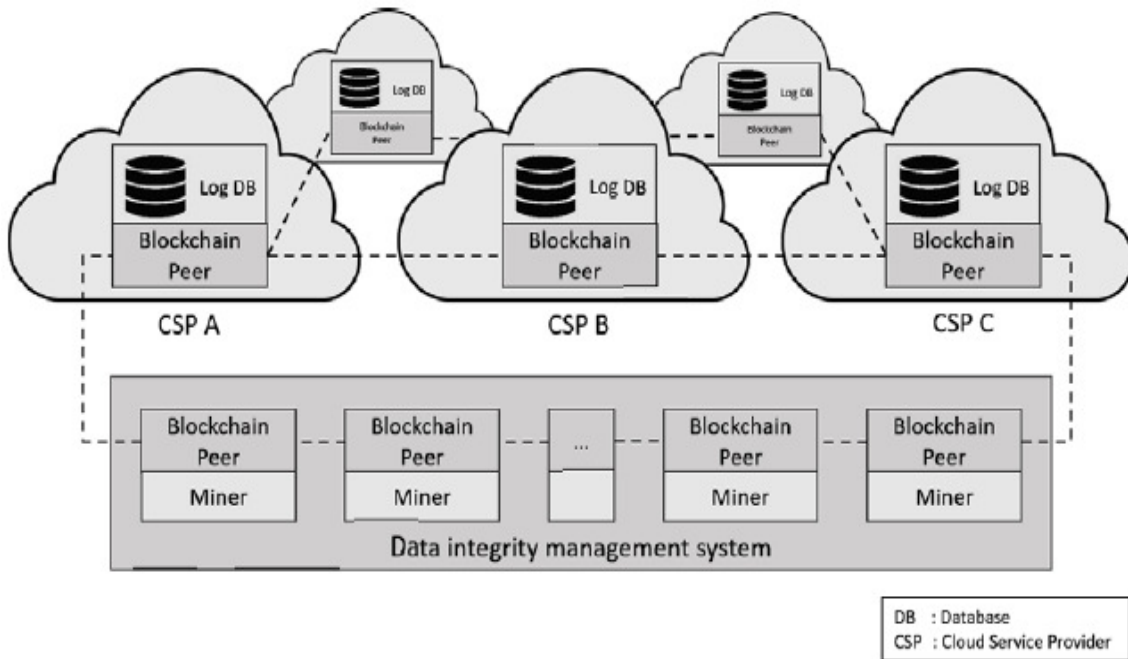


Figure 1. Blockchain based data integrity management system

Figure 1 shows the structure of the blockchain based data integrity management system. In this paper, the data of CSP is stored by itself, but the procedure for verifying the integrity of the data is performed through blockchain. Data that requires storage for cloud forensics is determined through agreement between CSP and CSC. At this time, the CSC should consider the additional cost incurred to store the data in the cloud environment. The collected log data is converted into a hash value through a hash function. These data are used to create a hash tree and construct a block. In the case of permission-less blockchain such as Bitcoin, all peers participating in the network can perform mining to create new blocks. However, this is not suitable for proposed system because not all CSP peers can be trusted. Also, if all CSPs participate in mining by PoW method, the proposed system is very inefficient because it needs to consume more computing power than mining power of CSPs. Therefore, only the data integrity management system performs making block and the each block consists of the hash value of the CSP data. One block can contain data of one CSP and the data of CSP participating in the system are stored in order. The generation period of the block is determined by the agreement of the CSPs participating in the system, and it is determined in consideration of the processing performance.

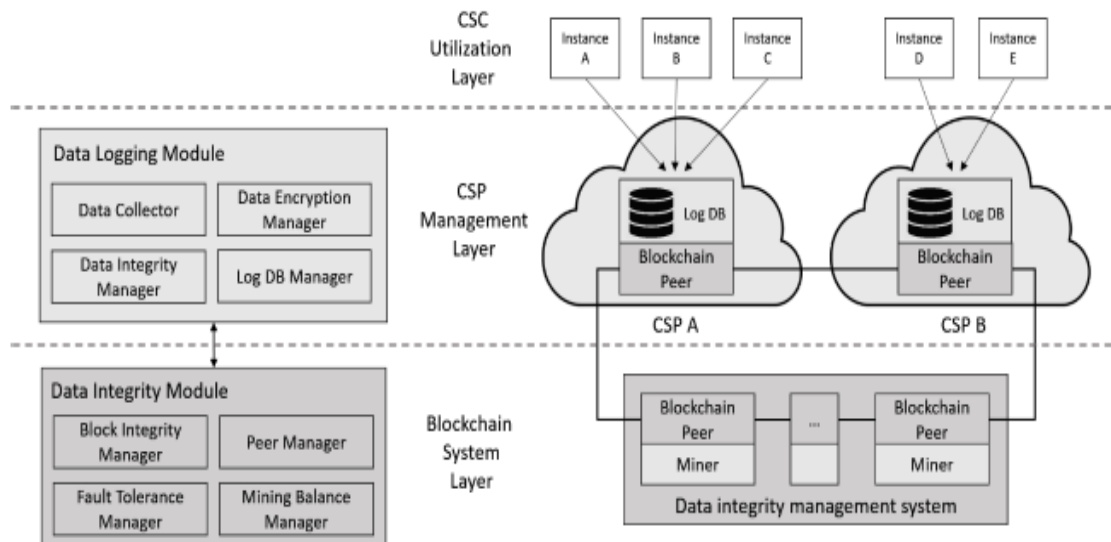


Figure 2. Overview and detailed functions of the proposed system

Figure 2 shows the overall flow and detailed functions of the proposed system. In CSC Utilization Layer, the data of instance used by CSC is stored in the CSP Management Layer, and the Blockchain System Layer manages the integrity of stored data. The Data Integrity Module consist of four functions: Block Integrity Manager, Peer Manager, Fault Tolerance Manager, and Mining Balance Manager. Details of each function are as follows.

3.1 Block Integrity Manger

The Block Integrity Manger performs integrity check on data received from CSP when new block is created. Integrity verification is the process of checking the hash value to see if the encrypted data has not changed and verifying that the data is being sent by that CSP.

3.2 Peer Manger

The Peer Manager monitors the number of CSPs participating in the network and adjusts the number of peers so that the Byzantine Generals Problem does not occur. When a block is created and propagated, it performs a task of maintaining a minimum number of $3n + 1$ peers so that n malicious CSP nodes do not interfere with the block generation process. In addition, the size of the block is adjusted to obtain the optimum performance considering the number of peers. It also manages the peers of the CSPs that want to join or leave the network.

3.3 Fault Tolerance Manger

The Fault Tolerance Manger performs tasks such as building a block by solving a fault situation such as branching or consensus falling into a deadlock when stacking blocks.

3.4 Mining Balance Manager

The Mining Balance Manager performs the task of adjusting the cycle of generating block time. If the period is not constant, size of the data in each block may be unbalanced, which may complicate the integrity verification of the data when doing cloud forensics. Therefore, by adjusting the minimum time and maximum time range in which a new block is created, it is possible to stack data of a proper size into one block.

The Data Logging module in CSP Management Layer consist of four functions: Data Collector, Data Encryption Manager, Data Integrity Manager, Log DB Manager Details of each function are as follows.

3.5 Data Collector

The Data Collector performs the task of collecting the service data or log that the CSC requested to be collected. It is recommended that you use a public tool that can be used for cloud forensics when collecting, for example snort to store network packet data.

3.6 Data Encryption Manger

Because data in the cloud environment may be related to the privacy of the CSC, the Data Encryption Manager performs encryption of the data collected by the data collector. Encrypted data is recommended to be encrypted using the CSC's public key.

3.7 Data Integrity Manager

The Data Integrity Manager manages the integrity of data collection and storage. It is difficult to trust CSP's integrity of data in an environment provided by CSP itself. This means that CSP manages the data that is stored before it can be used as evidence for the cloud forensic investigator by this procedure.

3.8 Log DB Manager

The The Log DB Manager performs the task of storing the collected data. The stored data is transmitted to the Data Integrity Management System after performing a hash operation.

4. PERFORMANCE CALCULATION

In this section, we compare transactions per second(tps) with the other cryptocurrency mechanism with our proposed system. One of the reasons why mining-based permission-less cryptocurrency are not used in other area is that the number of transactions per second is too small. The tps formula of cryptocurrency is as follows.

$$tps = \frac{Blocksize}{Blocktime \times Size\ of\ the\ Transaction} \quad (1)$$

VISA, a credit card company, handles 100,000 transactions per minute in 2016[15]. Compared to that, the tps of the cryptocurrency is too low. For example, 6.41tps for Bitcoin, 15.65tps for Ethereum, 26.67tps for Zcash (unshielded), and 6.67tps for Zcash (shielded). The contents are shown in Table 1.

Table 1. Comparison of TPS of the proposed system with other cryptocurrency

| | Blocksize (MB) | Blocktime (sec) | Transaction size (bytes) | tps |
|--|---------------------------|----------------------------|-------------------------------------|------------|
| Bitcoin | 1 | 600 | 260 | 6.41 |
| Ethereum | 4.7 | 14.3 | 21000 | 15.65 |
| Zcash(unshielded) | 2 | 150 | 500 | 26.67 |
| Zcash(shielded) | 2 | 150 | 2000 | 6.67 |
| Proposed System (tps : per CSP) | 3.2 | 600 | 32 | 166.67 |

The proposed system, assuming that uses a permission blockchain such as Hyperledger[16], the manager can revise the chaincode to set the rule. In the proposed system, when the data of one CSP is stored in a cycle of 10 minutes, the transaction of the blockchain system in 10 CSP environment can be thought to occur once a minute. To assume the size of the log data, previous research about security data logging system for cloud forensics proposed by Zawoad et al.[17] each log uses SHA-256 hash function. Therefore, it is assumed that our proposed system also uses that method. Assuming that the log is generated once per second, about 600 logs are created because one block is stored every 10 minutes in one CSP. The hash value of each log can be defined as transaction. If the size of encrypted log is 100byte, size of one block can be 3.2MB including Hyperledger's block header in order to save this log in hash tree. Based on this situation, we calculate about 1667 tps of the proposed method and about 167 tps per CSP because there are 10 CSPs. The graph comparing TPS with other cryptocurrency is shown in Figure 3.

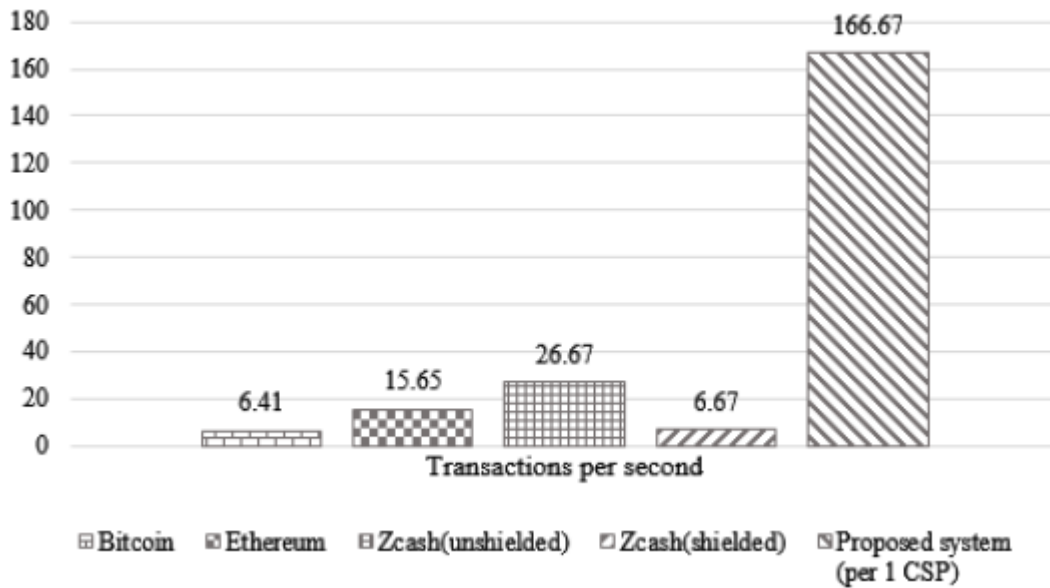


Figure 3. Compare transaction per second with other cryptocurrency

In Figure 4, we check tps according to the number of CSP participating in the proposed system. The horizontal axis represents the number of CSP. In the proposed system, the blocktime and the transaction size of the block chain are assumed to be the same, so the overall TPS shows an increasing tendency. The tps per CSP shows a certain range depending on the number of CSP. This is because the hash tree is organized in a binary tree. If the number of logs is less than the available number of hash tree, the height of the hash tree is expanded. It can be seen that the tps per CSP decreases as the number of logs and the size of the hash tree become equal.

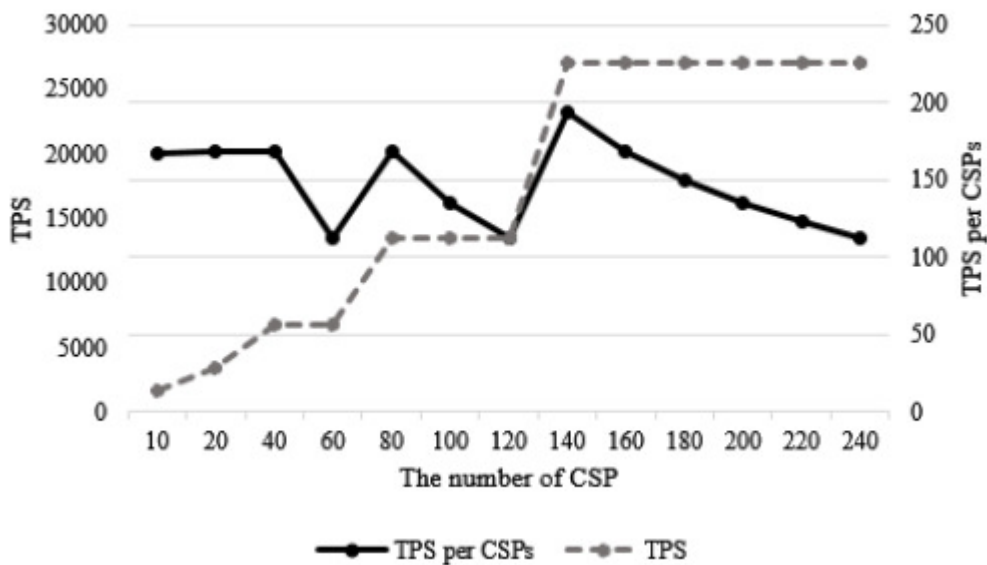


Figure 4. Comparison of tps according to the number of CSP

Permission blockchain such as Hyperledger cannot guarantee the accuracy of the tps because the user can arbitrarily control the size of the block and the block generation period, and the performance may vary depending on how the environment is configured. In addition, the proposed system does not consider the time required for consensus and the time required to process each transaction, so actual performance is expected to be lower. However, the existing PoW-based permission-less blockchain network has a low processing speed because untrusted persons are allowed to participate in the network while maintaining reliability. Thus, the processing speed of a permission block chain can be expected to be faster.

5. CONCLUSIONS

In this paper, we investigate the reason for logging in cloud environment for cloud forensics and propose the permission blockchain based data integrity management system. The proposed system is able to guarantee the integrity of data while processing more transactions than existing permission-less based blockchains. However, there is a limitation that the performance evaluation of the present system can not perform the actual evaluation merely by comparing the calculated result values by calculating the expected data size. The proposed system can be used as one of the methodologies for coping with security incidents in the cloud environment. As future work we collect network data with snort and perform simulation to calculate accurate tps by using Hyperledger. The reason for choosing network data is that cloud environment has a complex network environment due to the virtual network configuration, and there are many incidents that exploit its vulnerabilities. We will also perform a performance evaluation comparing the time required for various consensus algorithms for comparison between permission blockchains.

REFERENCES

- [1] K. Kent, S. Chevalier, T. Grance, and H. Dang, "Guide to integrating forensic techniques into incident response," NIST Special Publication, pp. 80-86, 2006.
- [2] Josiah Dykstra, Alan T. Sherman, "Acquiring forensic evidence from infrastructure-as-a-service cloud computing: Exploring and evaluating tools, trust, and techniques", Digital Investigation 9, 2012
- [3] J. Dykstra and A. T. Sherman, "Understanding issues in cloud forensics: two hypothetical case studies," in Proceedings of the Conference on Digital Forensics, Security and Law, 2011, p. 45.
- [4] K. Kent, S. Chevalier, T. Grance, and H. Dang, "Guide to integrating forensic techniques into incident response," NIST Special Publication, pp. 80-86, 2006.
- [5] S. Zawoad and R. Hasan, "Digital forensics in the cloud," DTIC Document, 2013.
- [6] Chun, Byung-Gon, et al. "Efficient Replica Maintenance for Distributed Storage Systems." NSDI. Vol. 6. 2006.
- [7] Nepal, Surya, et al. "DIaaS: Data integrity as a service in the cloud." Cloud Computing (CLOUD), 2011 IEEE International Conference on. IEEE, 2011.
- [8] Gaetani, Edoardo, et al. "Blockchain-Based Database to Ensure Data Integrity in Cloud Computing Environments." ITASEC. 2017.
- [9] Nakamoto, Satoshi. "Bitcoin: A peer-to-peer electronic cash system." (2008): 28.

- [10] King, Sunny, and Scott Nadal. "Ppcoin: Peer-to-peer crypto-currency with proof-of-stake." self-published paper, August 19 (2012).
- [11] Kishigami, Junichi, et al. "The blockchain-based digital content distribution system." Big Data and Cloud Computing (BDCloud), 2015 IEEE Fifth International Conference on. IEEE, 2015.
- [12] Azaria, Asaph, et al. "Medrec: Using blockchain for medical data access and permission management." Open and Big Data (OBD), International Conference on. IEEE, 2016.
- [13] Oudejans, Joris, and Zekeriya Erkin. "DecReg: A Framework for Preventing Double-Financing using Blockchain Technology." Proceedings of the ACM Workshop on Blockchain, Cryptocurrencies and Contracts. ACM, 2017.
- [14] Christidis, Konstantinos, and Michael Devetsikiotis. "Blockchains and smart contracts for the internet of things." IEEE Access 4 (2016): 2292-2303.
- [15] Jan Vermeulen, "VisaNet – handling 100,000 transactions per minute" [Online]. Available: <https://mybroadband.co.za/news/security/190348-visanet-handling-100000-transactions-per-minute.html>. 2016.12.
- [16] Cachin, Christian. "Architecture of the Hyperledger blockchain fabric." Workshop on Distributed Cryptocurrencies and Consensus Ledgers. 2016.
- [17] Zawoad, Shams, Amit Kumar Dutta, and Ragib Hasan. "SecLaaS: secure logging-as-a-service for cloud forensics." Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security. ACM, 2013.

AUTHORS

Jun Hak Park

Jun Hak Park received B.S. degree in Department of Computer Science and Engineering from Kyung-Hee University, South Korea, in 2016. He is pursuing his M.S. degree in the Department of Computer Science and Engineering at Kyung-Hee University, South Korea. His research interests include Cloud Computing, Cloud Forensics, and Blockchain.



Jun Young Park

Jun-Young Park received his B.Eng. degree in Computer Engineering from Hannam University, Korea, in 2010, and a Master's degree in Computer Engineering from the Kyung Hee University, Korea in 2012. He is currently working toward a Ph.D. degree in the Department of Computer Science and Engineering at Kyung Hee University, Korea. His research interests include cloud computing, mobile cloud computing, cloud computing security, security-as-a-service.



Eui-Nam Huh

Eui-Nam Huh earned a B.S. degree from Busan National University in Korea, a Master's degree in Computer Science from the University of Texas, USA in 1995, and a Ph.D. degree from the Ohio University, USA in 2002. He is the director of Real-time Mobile Cloud Research Center. He is a chair of Cloud/BigData Special Technical Committee for Telecommunications Technology Association(TTA), and a Korean national standards body of ITUT SG13 and ISO/IEC SC38. He was also an Assistant Professor at Sahmyook University and Seoul Women's University, South Korea. He is now a Professor in the Department of Computer Science and Engineering, Kyung Hee University, South Korea. His research interests include Cloud Computing, Screen Contents Coding(Cloud Streaming), Internet of Things, Distributed Real-Time Systems, Security, and Big Data



INTENTIONAL BLANK

AN INNOVATIVE SOCIAL MOBILE PLATFORM TO SUPPORT REAL-TIME COMMUNICATION IN PEER TUTORING

Meghan Wang¹ and Yu Sun²

¹Valencia High School, Placentia, USA

²California State Polytechnic University, Pomona, USA

ABSTRACT

This paper looks into the reasoning, context, and process behind the creation of the Step Up App to be used in the Step Up Club. The Step Up Club is a peer tutoring high school organization that allows students to tutor one another. The paper explains the background and the issues that exist with peer tutoring regarding challenges in communication between tutors and tutees. The app aims to provide the solution for many of those problems in creating a new platform in which students can communicate to one another about any questions regarding school academics.

KEYWORDS

Computer Science, Communication, Peer Tutoring App, Academic Help, Mobile Computing

1. INTRODUCTION

Peer tutoring is a method in which people who hope to learn help each other by teaching one another [1]. This less common form of tutoring could bring many advantages to learners and is being studied to see the advantages it creates [2]. In fact, the area of peer tutoring is being researched from all areas including not only educational but also social and psychological facets [3]. In recent years, high schools secondary education has also been a large area of discussion in investigating different areas of schools as well as changing existing curriculums to analyse students' performance [4]. Peer tutoring within high school students has already shown positive results [5].

This research paper explains how the Step Up App was created in order to help in peer tutoring communication that was absent in the organization before. The Step Up App is a peer tutoring App allowing students to message peers about questions concerning school or homework. It consists of a real-time chat, allowing students to join a large discussion room, or to message tutors on a one to one basis. The app also contains a leader board, from students ratings of tutors, allowing other students to evaluate how helpful certain tutors are.

High school students often face much stress and need help on their school work. School clubs and organizations as well as the Parent Teacher Association have all made efforts to provide tutoring or help in standardized testing for students. However, students often are unwilling or unable to afford tutoring and also may not feel that tutors who have not experienced the actual course can understand their problems.

The Step Up Club is a high school organization at Valencia High School, created to help students receive one on one peer tutoring from other high school students for free. The idea of peer tutoring was implemented in order to allow students to receive help from other students who have had first hand experience in taking a class and mastering it or excelling in the course. Tutors for each of the subjects are chosen after submitting their grades in the course they have already completed.

However, the organization only provided help when students and tutors met up with one another in person. If students realized they needed help once they left the school's library where they would meet with tutors, it would be difficult to receive the help they needed. Other circumstances such as scheduling to meet with tutors in person also create similar challenges with a lack of other forms of communication.

The Step Up App creates a platform to solve this issue by allowing students to immediately communicate with tutors even at home, using the real-time chat. The leader board, displaying top ranked tutors are recognized and awarded by the club, providing an incentive for students to help others. The app allows a more convenient form of communication for both tutors and tutees.

This paper is organized as follows. Section 2 discusses the challenges that the Step Up organization faced which eventually led to the creation of the app. Section 3 discusses the solution the app has to offer to the challenges. Section 4 approaches how each solution works to solve the challenges. Section 5 discusses some other similar ideas to the Step Up App, and Section 6 concludes the research paper.

2. CHALLENGES

The examples with many students, both tutors and tutees, led to a closer analysis at a pattern seen throughout of challenges the Step Up organization faced.

2.1. Inconveniency

One of the many challenges faced by the Step Up organization was the inconveniency of requiring tutors to meet face to face with tutees. As high school students, both parties face many tight schedules from after school competitions such as decathlon to competing in seasonal sports such as volleyball. Having to schedule sessions with peers became a monotonous process as both sides would repeatedly have conflicts in timing. This was one of the many inconveniences of the structure of the organization.

Additionally, having tutors and tutees meeting at school had an even narrower time frame due to the opening and closing hours of the high school's library, where students could meet and have textbooks available to borrow. This time frame lasted from when school ended at 2:45 to 4:00. The time frame is not only short, but leaves students very little amount of time in being able to schedule a tutoring session. These inconveniences led to much frustration with the students and an unwillingness to schedule sessions. Many students who try to meet with one another outside of school often have difficulties as well in attempting to find a proper location where both sides are easily able to meet at the same time.

2.2. Lack of Motivation

From the beginning of when the Step Up organization to present time, there has been a great trend in decreasing numbers of both tutors and tutees participating as a member of the club. When members first joined, there was a large group in both sides of the club, as this innovative organization was formed at the school. Students had higher participation rates as well as interest

in the program. As the high school students began to move up in grade level, taking on more challenging courses and extracurricular activities, a fall in interest began showing due to the inconvenience in having to stay for complete sessions and scheduling with tutors.

Over time as the Step Up organization grew older, students began developing a lack of motivation in participation as both a tutor and tutee. A portion of the growth in lack of motivation throughout the club was due to the frustration and tenuous process of attempting to communicate between both parties in scheduling tutoring session. Throughout the year, tutoring sessions would diminish as students no longer wanted to take time out to schedule a session for simple questions concerning school work.

Furthermore, another contribution to the lack of motivation seen in participants of the Step Up organization was largely due to the fact of the creation of a competing organization at Valencia High School known as “Tiger Tutoring,” that branched off of a larger organization known as National Honor Society. National Honor Society, or NHS, is a much larger nationwide organization for high schools that provides an incentive of recognition after student’s participation in tutoring other students. NHS tutors would receive recognition through honor cords and a cap tassel for graduation. This newly created organization not only drew away participants from Step Up, but also created a new competition of tutoring to Step Up. It provided similar hours of community service hours to students and also brought in more incentives of being recognized for participating.

2.3. Timing

Issues concerning timing became a large challenge generally for tutors. One of the main issues with Step Up’s original structure is its rigid form of communication and restricted person to person interactions. The inconveniency with scheduling sessions is not the only issue. A larger issue holds in that tutees often realize they need the most help in certain facets upon returning home and working on problems on their own. At these moments when tutees on their own, they need the help of tutors but are unable to because the tutors are only available through actual meetings.

Scheduled tutoring sessions also incur timing issues with certain deadlines that teachers have set for students. Because of the packed schedules every student takes on, there may be occurrences where tutors and tutees can not decide on a meeting time to meet in person before the deadline of a large project or a test date.

2.4. Challenge faced by Specific Student

In the year 2015-2016, one female student facing difficulties in understanding the content of her chemistry course was recommended by her teacher in receiving help from a tutor at Step Up Club. Over the semester, the tutor and tutee would meet up after school, in the library at school to receive help. However, the tutee mentioned numerous times that she often struggled with her work at home even after receiving help at school. The tutee had many questions that she had not thought of to ask the tutor until returning home and reviewing the topics of the unit.

3. SOLUTION

The problem surrounding many of the challenges with the original Step Up organization lies in the rigid structure of the club. Through the app itself along with a leader board and real-time chat room for discussion sessions, the Step Up App is created as the solution to many original problems.

Figure 1 shows an overview of the architecture of the app. The app is based on a typical server-client communication model. We have implemented the app in Android, while the backend is supported by Google Firebase. The reason to choose Firebase is its enhanced support on real-time data communication and synchronization.



Figure 1. The Architecture of the Step Up App

| US Smartphone User Penetration, by Age, 2014-2020 | | | | | | | |
|--|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| <i>% of mobile phone users in each group</i> | | | | | | | |
| | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 |
| 0-11 | 28.7% | 35.3% | 41.0% | 45.0% | 47.1% | 48.9% | 49.7% |
| 12-17 | 71.0% | 78.5% | 84.0% | 89.0% | 91.0% | 92.0% | 92.9% |
| 18-24 | 85.2% | 90.1% | 94.9% | 98.0% | 98.4% | 99.0% | 99.4% |
| 25-34 | 84.0% | 90.2% | 95.2% | 97.0% | 97.1% | 97.2% | 97.3% |
| 35-44 | 81.6% | 87.7% | 92.0% | 94.0% | 94.8% | 96.1% | 97.0% |
| 45-54 | 66.9% | 75.8% | 82.8% | 88.3% | 92.8% | 95.9% | 97.9% |
| 55-64 | 59.1% | 67.6% | 75.7% | 80.9% | 85.5% | 89.5% | 93.0% |
| 65+ | 36.6% | 40.7% | 44.4% | 49.6% | 52.8% | 55.7% | 55.4% |
| Total | 67.6% | 73.8% | 79.0% | 82.7% | 84.8% | 86.5% | 87.3% |

Note: individuals who own at least one smartphone and use the smartphone(s) at least once per month
 Source: eMarketer, Feb 2016

204491 www.eMarketer.com

Figure 2. eMarketer's Table of US Smartphone User Penetration [6]

3.1. Holistic App

The design of the app itself has the main goal of portability and ease of usage. The original issue was due to the inflexible composition of the club in having only person to person real life meetings. The app offers a real-time chat that allows tutors and tutees to maintain and create tutoring sessions even outside scheduled meetings.

The rationale behind the creation of the app was to create a system that could be used anywhere and anytime, to increase an efficiency and convenience for all members of the club. The idea of an app as the larger goal to attack the challenges emerged from noticing the everyday usage of people on apps. Almost all high school students have smart phones that they carry around on a daily basis. Creating an app that can be easily downloaded on these hand held devices becomes a much easier process for all students rather than having to schedule and meet someone in person. eMarketer has in fact studied teenagers usage of smart devices as seen in the chart below, and the

trend has only shown an increase in the percentage of the teenage age group [6]. The trend even extends to 2020. Implementing an educational app for teenagers to use may also add a positive impact to the large usage of young people in mobile devices.

3.2. Leader Board

The leader board is a feature in the app that consists of a rating system for students on each tutor. The rationale behind creating such a screen within the app was to offer an incentive to students. As seen in the competition of the tutoring organization known as Tiger Tutoring created by the National Honor Society which offered a recognition of being a tutor, the system of a Leader Board similarly offers a recognition. In fact the leader board offers a more competitive form of recognition because the ratings between every tutor becomes an incentive that makes members of Step Up all strive to become better tutors for one another.

The rationale behind the leader board was not only to create this form of motivation but also in bringing in an element of reliability for students. In having such a rating system, students become more aware and are able to differentiate from the better and worse tutors. This creates a much more trustworthy system that provides the best form of resources for the students who are tutees needing help in a particular area.

3.3. Real-Time Chat

The main feature and idea behind creating the Step Up App was to implement a real-time chat that allows for the flexibility in time between tutors and tutees. The rationale behind the app was its ease in being used between tutors outside of scheduled tutoring sessions. It not only provides a connection between original tutors and tutees that continue meeting both in person and through messaging, but it may also induce new participants because of its flexibility and lack of binding between a tutor and tutee. Additionally, the rationale behind the real-time chat between two people also led to a larger group discussion that included all members. The idea behind the larger group discussion was to promote a flexibility in timing of allowing all people to answer in urgent need. However, the basic idea of one to one tutoring that serves as the foundation of the Step Up club is still maintained.

The real-time chat is implemented using the data service – Firebase [7] hosted in Google Cloud [8]. Firebase is a popular mobile backend as a service solution to support real-time data synchronization and communication across multiple clients. The messages that are sent within the chats on the app are all stored through the online Firebase database. Individual chat rooms between a tutor and tutee is stored under the tag “Tutor Messages,” stored by time order as value under the tag of the user’s name. The larger chat room with all users is identified under the tag of “Messages” and the values are in order the time in which the messages were sent as well.

4. METHODOLOGY OF SOLUTION/EMPIRICAL RESULTS

The Step Up App is a new edition to the Step Up organization that provides numerous elements and methods in helping solve many of the issues faced including inconveniency, lack of motivation, along with timing. The app also aims to create a new feature to the organization as a whole. The methodology in this section serves to explain in detail each of the specific features of the app that serve as solutions to particular challenges.

4.1. Solving Inconveniency

The Step Up App helps in solving many of the issues surrounding the inconveniency of the rigid original structure of the organization. The app itself is the main solution to the inconveniency

through its portable method of communication between tutors and tutees. With the publication of the app, the two parties no longer have to schedule person to person meetings and rather have a new means to communicate with each other whenever and wherever. The real-time chat allows for convenience between tutors and tutees in much more efficiently being able to communicate with one another regardless of each sides' schedules.

In the situation where tutors and tutees schedules conflict, the tutee could easily first ask the tutor a question using the app. Once the tutor has the time to respond, they could quickly respond to one another. Students could also use the app to potentially schedule a real life meeting in person with their tutors if they hope to do so. The app provides a new platform that still permits the original basic composition and purpose of the Step Up organization, but adds along a new highlight to the club. Students who still prefer the original person to person contact still have the more intimate method available. At the same time, if needed, they have the app to remain in contact with tutors at home. During busier testing seasons, both tutors and tutees have a much more convenient method of communication through the app. Students who face trouble in finding methods to meet up outside of school similarly can use the app as a form of communication.

The app itself is also incredibly easy to move from screen to screen, whether signing in, creating a new account, or chatting with tutors. People can easily install the app on their phones from the Play Store, create an account, and begin interacting with tutors. Each of the features of the app can be selected from the "Home Page" that users immediately see upon signing into the app. When students select particular tutors, they are able to chat with them one on one. Otherwise, students can also join the large chat room that hold all registered members of Step Up and have discussions of their topic.

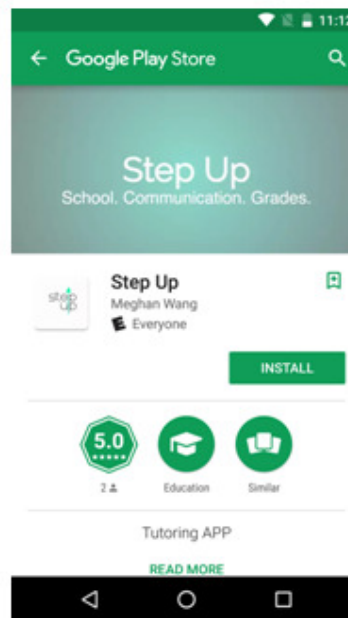


Figure 3. Screenshot From Smart Phone of Google Play Store

4.2. Creating Motivation

Another large issue the Step Up organization faced was in the lack of motivation from both tutors and tutees over time. The majority of this issue is solved with the publication of the app yet again. A decrease in motivation was largely seen due to the lack of convenience that previously existed

when tutors and tutees had to go through long processes of scheduling with one another repeatedly and each facing differences in timing that eventually led to frustration and irritation, causing the lack of motivation. The app helps in solving the issue of the inconvenience.

The leader board portion of the app is also an important component that helps in creating the motivation necessary for the success of the organization. The leader board element of the Step Up App is a screen that ranks the top ten best tutors, as seen ranked by other students who have received help from those tutors. Students are able to rank each tutor on their profile pictures with a rating of five being “extremely helpful” to one being “not very helpful.” The leader boards uses a function made during the creation that averages the rankings of each tutor and lists them out. At the conclusion of each month, the top ranked in the leader will be recognized for their help and dedication to the organization.

One of the other factors to the fall of motivation in members of the Step Up club can mainly be attributed to the existence of the competition from the other organization of tutoring created at the school by the National Honor Society. The addition of the Step Up App helps diminish much of the lack of motivation in participating in Step Up because of the new feature that has been created that the competition lacks. It creates a new highlight that will draw members into because of its innovative aspect.

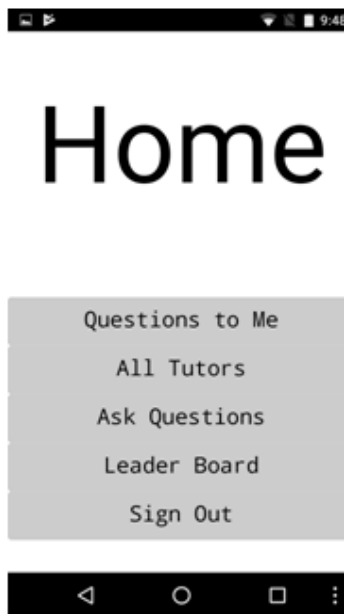


Figure 4. Home Page of Step Up App

4.3. Establishing Much More Flexible Timing

Timing becomes much more flexible and no longer is a challenge because of the Step Up App. Using the app, tutors and tutees can easily communicate with one another from anywhere at anytime. The rigid scheduling issue is solved with this flexibility. Students having various schedules no longer need to face the original issue. The challenge surrounding timing of how students having questions after tutoring sessions is thus also solved. When tutees are at home, working and realize there is information they do not understand and did not originally imagine it as a difficulty, can easily contact their tutor with ease through the app.

Additionally, if students face challenges on their own and ask their tutor, but the tutor missed the message, forgot to respond, or is busy, the issue with tight deadlines may seem to arise again. However, students are always able to join the large chat room and discuss with all other students or they have the ability to contact all other tutors one-to-one by selecting their profile and discussing if they prefer to do so in that manner. The flexible timing with the app helps resolve issues against the less flexible timing with school assignments, tests, and large projects by offering help and communication that is essentially unrestricted by timing.

5. RELATED WORK

The Step Up App provides a Peer-to-Peer Tutoring that some other Apps and online resources also contain similarities too and also offer similar services.

5.1. Sesh Mobile App

The Sesh mobile app was created by a company from Vanderbilt University and Stanford University students, known as Vanford, and the purpose was to allow students to request for tutoring sessions during any time. The company wanted to create a more expansive community of learning for students leading to the creation of the app and it started with only the two universities but hopes to expand. The app was used to request in-person sessions, also known as a “sesh” with tutors who have had to apply by uploading their transcripts of what classes they offer. Each tutor earns \$20 per hour [9].

This app is similar to the Step Up App in the hopes of creating a community where all students are able to communicate about academics in one area. However, the Sesh mobile app is used solely to request an in-person tutoring meeting. The goal of the Step Up App is to allow both in-person as well as easily accessible online communication to create efficiency for students. Additionally, Step Up has no requirement for tutors to submit an application because all members of Step Up are permitted to help others. While an application can assure reliability, the Step Up App is made reliable through a tutee rating system. Additionally, Step Up has no costs at all to receive the help students need. The Sesh App promotes an all day and night availability, but this is much more easily achieved through Step Up which does not require for either tutors or tutees to travel when the tutee needs help late at night.

5.2. Brainfuse

The Brainfuse online tutoring website offers help to students through a 24/7 online tutoring and writing lab system. It offers an online peer tutoring system that allows students to contact one another as they believe in peer tutoring as the method that often allows students to learn more easily. The “Online Learning Platform” of Brainfuse is aimed at helping colleges and universities in creating a peer-to-peer learning community [10].

The Step Up App has many similarities to Brainfuse in its core idea of a providing a community for students to peer tutor one another. While the Brainfuse website also provides an actual applied tutor as well, it is only focused on colleges and universities. The Step Up App is designed as a part of an organization in one high school, but aims at being user friendly to people in all levels of education. Additionally, the mobile App itself is effortlessly accessible with a click away on the phone that an online website does not provide for.

6. CONCLUSION AND FUTURE WORK

The Step Up App is an innovative approach in bringing in a peer tutoring service to a larger community. It provides a new feature to the Step Up organization in peer tutoring by solving the original challenges faced including inconvenience, lack of motivation, as well as inflexibility in timing. The Step Up App can be downloaded and installed from the Google Play Store onto mobile devices [11].

The Step Up App is able to dismiss the problems the club faced through its easily accessibility from being downloaded on the smart phone used commonly by high school students. The app itself is not only easy to download and install on smart devices but also has easy to use elements of signing up and logging in and each user having a home screen that contains the various features they may use. These features include but are not limited to “Questions to Me,” “Ask Questions,” and “Leader Board.” The “Ask Questions” tab leads to a large chat room that includes all members. This is a new idea implemented in the club which originally only targeted one on one tutoring. This large real-time chat room allows all students to quickly receive an answer for a question because all members are present in it to answer questions. The leader board was one of the main ideas behind the creation of the app in encouraging more motivation of members. It creates a platform in which all students are able to rate each other in the tutoring reliability and accuracy. This creates a much more trustworthy app that does not permit students from randomly answering others questions without actual knowledge. The leader board also creates an incentive that encourages students to answer others questions and become recognized for doing so.

In the future, there is still much to improve and update for the Step Up App. The Step Up App is still in the processes of developing a more modern styled user interface. The current user face includes each of the features and easy to access. However, a more modern styled user interface could please users in a more aesthetic manner. Additionally, the Step Up App is working on a possible limitation of users exploiting the rating system. Users may purposely vote positively for people they know to allow them to rank on the leader board. Step Up’s plan in trying to solve such a scenario is in the making of creating a monitoring system within the board of Step Up. Additionally, Step Up has in plan of trying a one month system of the leader board, which means for every one month, the rating will be cleared and all begin from zero again.

This form of App type that offers a peer-to-peer tutoring is on-demand and has not been largely invested yet but could become an extremely useful application not only within the high school but in all areas and levels of academics.

REFERENCES

- [1] Goodlad, Sinclair, and Beverley Hirst. *Peer Tutoring. A Guide to Learning by Teaching*. Nichols Publishing, PO Box 96, New York, NY 10024, 1989.
- [2] Topping, Keith J. "The effectiveness of peer tutoring in further and higher education: A typology and review of the literature." *Higher education* 32.3 (1996): 321-345.
- [3] Cohen, Jiska. "Theoretical considerations of peer tutoring." *Psychology in the Schools* 23.2 (1986): 175-186.
- [4] Boyer, Ernest L. *High school: A report on secondary education in America*. Harper & Row, Inc., 10 East 53rd Street, New York, NY 10022, 1983.

- [5] Fuchs, Lynn S., Douglas Fuchs, and Sarah Kazdan. "Effects of peer-assisted learning strategies on high school students with serious reading problems." *Remedial and Special Education* 20.5 (1999): 309-318.
- [6] "Teens' Ownership of Smartphones Has Surged - eMarketer", Emarketer.com, 2016. [Online]. Available:<https://www.emarketer.com/Article/Teens-Ownership-of-Smartphones-Has-Surged/1014161>. [Accessed: 01- Aug- 2017].
- [7] Link, Georg JP, et al. "Evaluating anchored discussion to foster creativity in online collaboration." *CYTED-RITOS International Workshop on Groupware*. Springer, Cham, 2015.
- [8] Moreno, Ismael Solis, et al. "An approach for characterizing workloads in google cloud to derive realistic resource utilization models." *Service Oriented System Engineering (SOSE), 2013 IEEE 7th International Symposium on*. IEEE, 2013.
- [9] J. Bolkan, "Sesh Mobile App Offers Peer-to-Peer Tutoring -- Campus Technology", *Campus Technology*, 2015. [Online]. Available: <https://campustechnology.com/articles/2015/02/04/sesh-mobile-app-offers-peer-to-peer-tutoring.aspx>. [Accessed: 07- Aug- 2017].
- [10] "Home - Brainfuse Online Tutoring", *Brainfuse Online Tutoring*, 2017. [Online]. Available: <http://home.brainfuse.com/>. [Accessed: 06- Aug- 2017].
- [11] "Step Up - Android Apps on Google Play." *Google Play*, Google, play.google.com/store/apps/details?id=appinventor.ai_MeghanHaoWang.Tutor.

AN INTELLIGENT SELF-ADAPTIVE SYSTEM TO AUTOMATE THE SPRINKLER CONTROL

Jiahao Li¹, Yu Sun², Fangyan Zhang³

¹Northwood High School, Irvine, CA, USA

²Department of Computer Science,
California State Polytechnic University, Pomona, CA, USA

³Department of Computer Science and Engineering,
Mississippi State University

ABSTRACT

It has been seven years since California is in serious drought. The dam holds rare water, and for some area the plants and people are suffered. While the technologies of desalination and reusing water is improving, it is significant if we solve the problem from the root, which is reducing water usage and saving water. Since eighty percent of water in California is used for agriculture and greening, it is efficient if we break through the system of irrigation. Currently, there are many ways to reduce watering in agriculture such as dropping water drops from pipes instead of spraying water; however, there are now resolution addressing the system of private watering yard in communities. The sprinkler device that we designed can contribute to reduce the water that is sprayed through sprinkler by adjusting the status of sprinklers (turning on or turning off) base on real-time weather conditions (temperature and soil humidity). Our purpose is to reduce the spraying water as much as possible if the weather condition allowed.

KEYWORDS

Sprinkler, Smart Control, Mobile App, Water Saving

1. INTRODUCTION

According to United States Geological Survey (USGS) [5], “As of May 23, 2017, the National Drought Mitigation Center estimates approximately 10.3 million people in California are currently affected by the drought.” In fact, California has been in drought for many years and the condition is getting worse and worse. Figure 1 shows that in recent record, still most of area in California suffers from shortage of water especially large city such as Los Angeles. Based on the current condition of drought, it is necessary for every resident of California to save water. With the rapid development of computer science and particularly Internet-Of-Things [6], it is possible to use technology to solve this problem in practice. Based on the fact that a lot of residents have to adjust the time of watering in our backyard very based on weather frequently [7][8], we decided to think about saving water resources though reducing unnecessary watering in every of these situations. In order to achieve the goal of reducing watering while keeping the plants healthy, we have designed and developed a sprinkler system that can adapt to the actual environment and be capable of turning on and off automatically according to the temperature and humidity in soil. Using this device, people do not need to worry about wasting water every time – the sprinkler can automatically turn on or off the sprinkler system based on the real need.

At present, people rarely pay attention to sprinkler all the time and they will not water the grass by hands, they simply set up the timer so the sprinklers can spray water automatically during certain period of the time. However, during this process, huge amount of water is wasted due to unnecessary watering in cool wet autumn evening. Thus, it is crucial to build an irrigation system that can save the whole process for people.

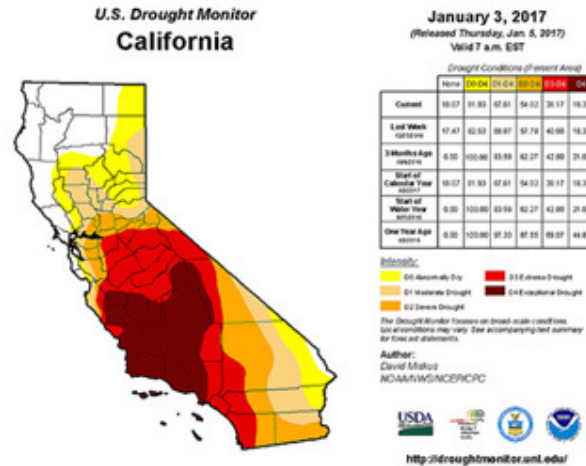


Figure 1: The U.S. Drought Monitor of drought situation of California in 2017 (adapted from [9])

To solve the issue, we have developed an intelligent smart irrigation system that contains two major components: 1) the Internet-Of-Things (IoT) system that uses temperature and humidity sensors to detect the actual soil environment data and send the real-time data back to the controller, where the smart decision will be made based on the data received; 2) A mobile app, “Servant Sprinkler”, would receive the real-time data and present them on the screen, which is easier for users to monitor the temperature and humidity since they can also turn on or off the sprinkler by hands.

The rest of the paper is organized as follows: Section 2 gives the details on the challenges that we met during the experiment and designing the sample; Section 3 focuses on the details of our solutions corresponding to the challenges that we mentioned in Section 2; Section 4 presents the relevant details about the experiment we did, following by presenting the related work in Section 5. Finally, Section 6 gives the conclusion remarks, as well as pointing out the future work of this project.

2. CHALLENGES

In this section, we discuss three challenges that confront us as we design the smart irrigation system. They are including accurate data collection, proper threshold setting, and effective interactions.

2.1. Challenge 1: Data Collection

The top one challenge is how to receive real-time data. If the operation of the device is totally depending on the weather report, it is useless. It is also impossible to collect data from several years and create a specific function to predict the future weather. Because there are different weather conditions in different areas and the sprinkler need to adjust itself base on weather condition, we need a device that can collect local environmental data constantly and present it to.

2.2. Challenge 2: Threshold Setting

The second challenge is that we needed to deal with how to set up the standards for both temperature and humidity, which are the values that sprinkler will shut down automatically if they are reached. Since the device mainly depends on the adjusting itself, it is crucial to set up relatively perfect standards to maximize the effort. If the standards are too high, the plants would die before the sprinklers are triggered to spray water; if the standards are too low, the situation of wasting water is still not solved. The research is needed for deciding the standards carefully.

2.3. Challenge 3: Interactions

The third challenge is how to make users interact with the device effectively and smoothly. As we all known, technology cannot replace human totally. There is probability that some tiny errors will occur, so human involvement is necessary. In this case, there might be some extreme weather that the temperature is low and the soil humidity is extremely low as well, which will not trigger the switch of sprinkler and the plants might suffer low temperature and poor moisture in the soil. This is the reason that the users need to know the real-time condition constantly. Since there is neither screen on the device nor a computer that is connected to it, it is necessary to visualize these codes to data, then to the sentence that people can read. Also, since it is not realistic for users to stay aside a computer to read the data, a moveable controller is needed which can show the meaning of data, temperature and humidity, and be manipulated by people at any time.

3. SMART SPRINKLER IN ACTION

To solve the above three challenges, we have developed a smart control device that can manipulate if there is water comes out or not automatically. As shown in Figure 2, the system contains two modules: the receiver that receives all data about humidity and temperature from sensors while is connecting to a Bluetooth launcher built with Arduino [11], which is connected to a Raspberry PI as the main controller [12]. The second module of the system is a power switch connecting to the main Raspberry PI controller. The controller is designed to revive the data through Firebase database, whose data is stored in Google Cloud. There is a program that we wrote in the Raspberry PI, which tells when to stop or release the water and spray to grass.

As it comes to collecting real-time data on temperature and humidity, one feasible solution is to send the data from sensors to cloud, which can be monitored through phone app. Since people may be out of home for most of the day time, it is nearly impossible to go to yard and keep paying attention of the data on sensor all the time, we develop an app that comes with the device. We upload the data from sensors to Firebase [10], then develop an app which can receive the data from computer and show it on the screen. Firebase is a cloud-based data synchronization and communication system that could be used to ease the data transfer across multiple devices and clients. Through this method, people are able to know the accurate temperature and humidity in their backyards.

On the other hand, a proper threshold to turn on or off the device has been tested and configured in order to enable the automated control. To accomplish that, we prepared two sensors for temperature and humidity; then wrote a program which gave orders to turn on the switch as long as the certain values are reached. Based on the special location of California, we set the limits as 24 Celsius degrees for temperature and 50% as moisture. Since the sunlight that strikes on the ground of California is relatively strong, the moisture is easier to evaporate, which is deal with by setting the standard for humidity as 50%. The highest average for Long Beach in 2010 was 29 °C, it is important for plants to receive sufficient water under high temperature and direct sunlight striking.

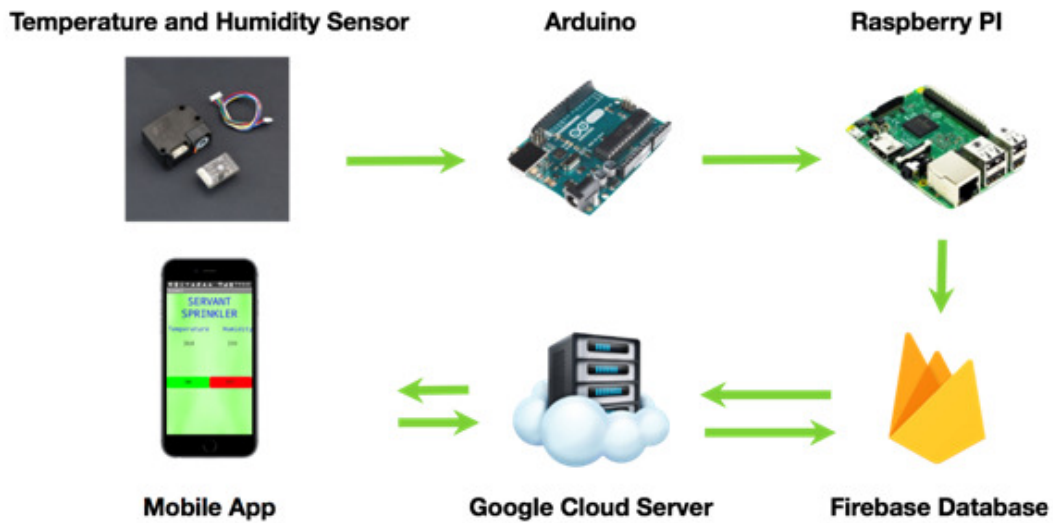


Figure 2. An Overview of the System

Finally, in order to address the challenge of how to interact with humans in case some extreme weather occurs, we decided to upload all the data from sensors to Firebase. There are already two sensors on the main board, and we added another element, which is a Bluetooth launcher. The data collected by the sensors will be launched to a cloud center call Firebase which could receive and store all the data. Then, a mobile app, Servant Sprinkler as shown in Figure 3, is developed for receiving the data from Firebase. In there, there are two blocks, one showing current temperature and the other one showing current soil humidity. This corresponding mobile app is a convenient tool for users to observe the real-time data. Besides that, there are two buttons, named “turn on” and “turn off”, can be used to activate or shut down the system by hands. If there are some kind of extreme weather with low temperature and low humidity, the users can manipulate the sprinkler and amount of water sprayed to grass base on their own wish instead of rely on technology only.



Figure 3. The Servant Sprinkler Mobile Controller App

4. EVALUATION OF SMART IRRIGATION SYSTEM

In order to accomplish this goal, the amount of water that is saved is the essential standard for evaluating the value of putting this product into market. Base on the observation, each neighborhood’s sprinklers sprays water for three minutes continuously. For average residential

sprinkler for lawn, each minute the water usage is about 1-6 gallons. Every day the sprinkler spray water at eight o'clock in the evening for three minutes, which tells that the sprinkler spray 54 gallons of water averagely. There are 150 houses in my community, which means that each day the water usage is 8100 gallons of water. However, based on the fact that there are approximately 20 percent of the days that the outside temperature is lower than 24 Celsius degrees and 7 percent that the humidity is higher than 50 percent as shown in Table 1, which means that over 30 days the sprinklers ignore the rain and spray the unnecessary water anyways, which is a total kind of waste. In my community, the spraying time is exactly three minutes at 8 pm every day, which is not reasonable because the cool weather will evaporate less water as it would during noon. After observation, each day one sprinkler in my yard sprays total of 5 gallons of water, which is the same for every sprinkler in my community. After connect the device to the sprinkler, my sprinkler sprayed total of 21 minutes per week per family, which means there is 35 gallons of water is used by one sprinkler. If the probability of weather lower than 24 Celsius degrees and humidity under 50 percent is 14 percent of time, there are 4.9 gallons of water is wasted in yard while there is a water shortage in California dam. After communicating with a Chinese factory, we made improvements about our device. We used their database as a basic stage for receiving and analyzing data. We further improved the device through designing and producing the packing of it and made it more like a mature product.

Table 1: The average Los Angeles temperature [14]

| High °F | Low °F | Month | High °C | Low °C |
|---------|--------|-----------|---------|--------|
| 68 | 48 | January | 20 | 9 |
| 69 | 49 | February | 20 | 10 |
| 70 | 51 | March | 21 | 11 |
| 73 | 54 | April | 23 | 12 |
| 75 | 57 | May | 24 | 14 |
| 78 | 60 | June | 26 | 16 |
| 83 | 64 | July | 28 | 18 |
| 84 | 64 | August | 29 | 18 |
| 83 | 63 | September | 28 | 17 |
| 79 | 59 | October | 26 | 15 |
| 73 | 52 | November | 23 | 11 |
| 68 | 48 | December | 20 | 9 |
| 75 | 56 | Year | 24 | 13 |

5. RELATED WORK

Sprinkler irrigation, distributing water by spraying it over the fields, has been applied for several decades. The water is sprayed from nozzles under the force of water pressure. [1] introduced sprinkler irrigation system and how to choose proper equipment. In order to deliver water to crops more effectively, there are several the study about sprinkler irrigation system. For example, [2] talked about how to formulate and solve mathematical expressions for the application depths and rates from a self-propelled, center-pivot sprinkler irrigation system. [3] proposed a method for evaluating the water application rate and uniformity coefficient of overlapping irrigation sprinklers. [4] conducted a research on relation between non-uniform sprinkler irrigation and crop yield. All these studies aimed to distributed water evenly then obtain better crop yield. None of

the studies involves saving water by improving sprinkler based the current temperature and humidity in soil.

6. CONCLUSION AND FUTURE WORK

In this paper, we present a practical solution to manipulate the switch automatically with a self-adaptive computer program that sets up the standards for the device to enable the water or not. As we can see from the device and the results of the experiment, the system effectively saves water everyday, offering an intelligent approach for users to manage the irrigation. As for the future work, the experimental design can be improved by replacing Bluetooth with other launcher since Bluetooth will be blocked by walls sometimes. Another direction to work on in the future is to enable sharing the data with the local community, so that a large dataset can be built. Using the dataset, models could be trained using machine learning techniques [13] which could be used to guide the all residents no matter whether they have the smart irrigation system or not.

REFERENCES

- [1] Pair, Claude H. "Sprinkler irrigation." (1970).
- [2] Heermann, Dale F., and Paul R. Hein. "Performance characteristics of self-propelled center-pivot sprinkler irrigation system." *Transactions of the ASAE* 11, no. 1 (1968): 11-0015.
- [3] Fukui, Y., K. Nakanishi, and S. Okamura. "Computer evaluation of sprinkler irrigation uniformity." *Irrigation Science* 2, no. 1 (1980): 23-32.
- [4] Stern, Jack, and Eshel Bresler. "Nonuniform sprinkler irrigation and crop yield." *Irrigation Science* 4, no. 1 (1983): 17-29.
- [5] USGS California Water Science Center, <https://ca.water.usgs.gov/data/drought/>
- [6] Xia, Feng, et al. "Internet of things." *International Journal of Communication Systems* 25.9 (2012): 1101.
- [7] Caswell, Margriet, and David Zilberman. "The choices of irrigation technologies in California." *American journal of agricultural economics* 67.2 (1985): 224-234.
- [8] Levy, Yvonne. "Pricing federal irrigation water: A California case study." *Economic Review Spr* (1982): 35-55.
- [9] U.S. Drought Monitor. <http://droughtmonitor.unl.edu/>
- [10] Google Cloud Firebase. <https://firebase.google.com/>
- [11] Faludi, Robert. *Building wireless sensor networks: with ZigBee, XBee, arduino, and processing.* " O'Reilly Media, Inc.", 2010.
- [12] Richardson, Matt, and Shawn Wallace. *Getting started with raspberry PI.* " O'Reilly Media, Inc.", 2012.
- [13] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *Journal of Machine Learning Research* 12.Oct (2011): 2825-2830.
- [14] U.S. Climate Data. <http://www.usclimatedata.com/climate/los-angeles/california/united-states/usca1339>

AUTHOR INDEX

- Abhinav Sharma* 127
Divya Sai Keerthi T 47
Dragan Ivanović 21
Eugène C. Ezin 69
Eui Nam Huh 127
Fangyan Zhang 171
Filippo Santarelli 57
Fréjus A. R. Gbaguidi 69
Georgia Kapitsaki 21
Helene Martin 115
Jean-Charles Pinoli 115
Jiahao Li 171
Jun Hak Park 127
Jun Young Park 127
Kostadin Kratchanov 01
Maruti Sairam Annaluru 81
Meghan Wang 161
Mekala Rama Rao 81
Mohammed A. Al-taha 139
Narayana Murthy BHVS 81
Ömer Mintemur 35
Pallapa Venkataram 47
Pierluigi Maponi 57
Pratap Reddy L 81
Ra'ad A. Muhajjar 139
Riccardo Piergallini 57
Ryan Alturki 97
Selma Boumerdassi 69
Sevil Sen 35
Solmaz Boroomandi Barati 115
Stephane Valette 115
Valerie Gay 97
Yann Gavet 115
Yu Sun 161, 171