International Journal of Computer Science and Application (IJCSA) Volume 3 Issue 2, May 2014 www.ij-csa.org
doi: 10.14355/ijcsa.2014.0302.01

A Theoretical Model for Detection of Advanced Persistent Threat in Networks and Systems Using a Finite Angular State Velocity Machine (FAST-VM)

Gregory Vert*[1], Bilal Gonen[2], Jayson Brown[3]

[1,3]Department of Computer Information Systems, Texas A&M University – Central Texas, Killeen, TX, U.S.A.
[2]Department of Computer Science, University of West Florida, Pensacola, FL, U.S.A.

*[1]greg.vert@ct.tamus.edu; [2]bgonen@uwf.edu; [3]jayson.brown@ct.tamus.edu

Received 25 November 2013; Accepted 15 January 2014; Published 15 May 2014

Abstract

Intrusion detection systems have undergone numerous years of study and yet a great deal of problems remain; primarily a high percentage of false alarms and abysmal detection rates. A new type of threat has emerged that of Advanced Persistent Threat. This type of attack is known for being sophisticated and slow moving over a long period of time and is found in networked systems. Such threats may be detected by evaluation of large numbers of state variables describing complex system operation and state transitions over time. Analysis of such large numbers of variables is computationally inefficient especially if it is meant to be done in real time. The paper develops a completely new theoretical model that appears to be able to distill high order state variable data sets down to the essence of analytic changes in a system with APT operating. The model is based on the computationally efficient use of integer vectors. This approach has the capability to analyze threat over time, and has potential to detect, predict and classify new threat as being similar to threat already detected. The model presented is highly theoretical at this point with some initial prototype work demonstrated and some initial performance data.

Keywords

Network Security; High Order Data Analysis; Intrusion Detection Systems; APT; State Machines

Introduction

Computer security efforts are engaged in an asymmetric fight against an enemy comprised of thousands of both independent and interrelated actors on an incredible number of fronts across a grand surface area. Given the asymmetry, it is inefficient and simply not practical anymore for analysts to manually investigate each new attack. While this is true for the general case of known attacks, it is especially true for Advanced Persistent Threats (APTs). APTs are a cyber-crime category directed at business and political targets (Michael K. Daly, 2009). Traditional intrusion detection requires processing large quantities of audit data, making it both computationally expensive and error-prone (Dell SecureWorks, 2012) (Karen Scarfone, 2012) (Y. Kareev, 2009) (Z. Othman, 2010). Limitation of traditional IDS (Intrusion Detection System) techniques are as much a function of the ability of a human to process large amounts of information simultaneously as they are limitations of the techniques themselves (Neil MacDonald, 2010). New approaches to pattern recognition that simplify the process for human beings could be beneficial in better analysis of attack data.

Background

Much research has been done in the area of intrusion detection but little in its application to APT. In some related works, Shuo looked at the application of variable fuzzy sets to reduce the probability rates of false alarms and increasingly low detection of threats (Shuo-Liangxun, 2011). Variable sets were used to create a dynamic model capable of adapting to an ever evolving system state. By defining multi-characters within a state, they can deploy system mechanisms that monitor network activity adaptively to improve

the implementation of a more reliable and more accurate set of detections. This in effect allows detection systems to be adaptive based on a multitude of personality traits found amongst complex systems rather than reactive to predefined security policies.

Vert et. al. looked at self organizing taxonomies that could detect and counter attack malware operations(G. Vert, 2012). In order to mitigate attacks on a system, methods were utilized to reduce sets of data containing attack signatures to only hardware specific state variables. By simply comparing previous states and how hardware state variables are affected over time, this research was able to build models that can respond to attacks based on fuzzy affinity matrices that can display natural trends of global system state variables over prolonged periods of data collection.

Work has also been conducted in the application of genetic algorithms (M. Hoque, 2012) to improve IDS operation. In this effort, GA was applied to detect previously undetected threat vectors, allowing more closely monitoring and conceptualization of network activity in a state system. Theories of evolution and ideas of adaptive characteristics based on these evolutionary state variables to dilute the complexity involved in threat detection were examined.

Very few approaches have sought to develop an analytic vector based algebra for detection. Some work however has been done looking at the role of visualization of security data (S. Eick, 1993) (Gensuo Han, 2012) (Bricken James, 1992) (Damballa, 2010). Previously, Vert et al. developed a conceptual mathematical model based on vector math and an analytic visual algebra for detecting intrusion attempts on computers. This approach was part of a larger project called “Spicule” (G. Vert, 1996) (R. Erbacher, 2002) (G. Vert, 1998). Spicule is a visualization technique that builds on this work and uses vectors to display the state of a system including its activity. This is a potentially powerful approach to intrusion detection because it uses vectors that offer themselves to various mathematical algebraic operations. Spicule can exist on many network hosts to aid in network security analysis.

This paper develops the theoretical model of a new concept referred to as a Finite State Angular Velocity Machine (FAST-VM) as a possible means to address APT on hosts in a network. It conjectures that APT could be detected and characterized by using very large data sets of network or host based state variables collected over long periods of time. The FAST-VM theoretical model has been developed to reduce high order state change data in networks and systems to the essence of what characterizes an APT presence in the system and to use a mathematical method that is computationally efficient and could analyze hundreds of thousands (>1,000,000) state variables in milliseconds.

Advanced Persistent Threats

By definition, an advanced, persistent threat, or APT, is a highly sophisticated networked entity, typical of organized groups that conduct hostile cyber-attacks against connected computers; whether on a local network or the internet (G. Vert, 2009). The intent of this type of threat may include one of the following (Dell Secure Works, 2012), i) gaining critical data found within government, financial or individual corporations, ii) maintaining access for future malicious intent, iii) manipulating data to disrupt the performance or operation of the target.

Approach

Mathematical Foundations

Integer based vector mathematics can be utilized to model state variables in a system. State variables are hardware and software attributes that change based on the systems operation; for instance CPU usage. State variables are indicators of the normal operation of a system and can assist in indicating the presence of abnormal operation or a threat.

The model proposed in this research is to use integer based vectors to model the state changes found in a system affected by APT presence. These vectors mitigate problems with the application of standard statistical methods for APT detection stemming from variation between hosts. Previous applications were extremely promising but focused on simpler exploits. Due to the subtle and sophisticated nature of APTs, there is no known upper bound on the number of state variable vectors that may be required for accurate detection. The underlying algebra has been shown experimentally to easily and efficiently handle thousands of state variable vectors without an appreciable degradation in execution time in large part due to being based on integer calculations. This allows for the reduction of high order state variable vector spaces to only the “needles in the haystack” which are indicative of APT presence.

Due to the persistent nature of an APT, vector modelling must also follow state variable changes over

a temporal range and across a geographic range as found in hosts on a network. Additionally, APTs are generally suspected to be mounted by sophisticated players. This necessitates the need to model a high order data set of state variables and their changes over time. The Finite Angular State Transition Velocity Machine (FAST-VM) has the capability to address these challenges. It can reduce the high order of state variable changes that have subtle changes in them over time to an easy to comprehend threat analysis that is the essence of APT's effects on a systems state variable.

A few types of mathematical approaches could be utilized to potentially detect APT presence. One of the most likely would be statistical methods. There are a few reasons why this approach is not taken in the FAST-VM model. The first of these is that APT attacks are systemic and complicated and the second is that statistical methods start to fail when hundreds of thousands of variable need to be analyzed in real time over a temporal range.

APTs are also expected to be detectable by way of monitoring a large number of networks and host state variables and analyze them simultaneously into a single aggregated picture for interpretation. The FAST-VM method has the capability to do this rapidly for any number of state variables at the same time. This could range from 10 to 1,000,000 state variables, all representing some aspect of a host's operation in a network of hosts under APT attack. Performance is discussed and analyzed in later sections of this paper.

Vector Mathematics Background

Vectors have a variety of expressions usually denoted by a lower case letter. They can have magnitudes and point to locations in space indicating a particular value for a state variable, or they point to a fixed location and grow in magnitude to indicate changes in state variable values. The FAST-VM uses a combination of these types of variables. The vector based approach of the FAST-VM model is simple computationally efficient as discussed later and lends itself to an algebra that can detect state variable changes or predict what a state variable will look like if an APT has affected its value. This allows the vectors to:

(i) **Detect** change in a state variable if an APT has infected a system.

(ii) **Predict** what a state variable vector would look like if an APT is present and affects its value.

Vectors in the FAST-VM model are given as $[x, y, z]$ = **v** where **v** is a changing vector based point in space representing the value of the state variable that the vector is modelling. Using this model it is possible to have hundreds of thousands of vectors modelling very complicated system operation in a network over a given time period. For example given a state variable vector **v** whose value at time t_0 is collected, and w for the same state variable collected at time t_1, the operation of subtraction has the following result if the values of the vectors have not changed:

$$\mathbf{w-v=y}$$

If **y**=0, no change to the state variable is detected suggesting no APT presence. This is referred as a **Zero** form. If **y**≠0, the effect of APT presence on the state is indicated and is referred to as an observer form or an Attack form if the APT was previously detected. If **w**≠**v** then **y**≠0 indicating the effect of APT on a given state variable. The jitter part of the model that is beyond the scope of this paper does address the notion of being able to say **w** and **v** are slightly different but essentially the same over time so as to reduce the FPR and FNR rates.

The above property is binary and the basis for detection of change in state variables affected by APTs over time. The vectors in the above example are referred to as the following:

- **v**: **Normal form**
- **w: Change form** – what has changed since v was sampled
- **y**: if **y**≠0, **w** − **v** is referred to as an **Attack form** and indicates APT presence. If **y**≠0 then no APT is present and **w** − **v** is referred to as a **Zero form**.

The algebra that predicts the effects of an unknown APT on what a state variable's value may be given a previously detected APT is defined as the following:

- **v: Normal form** – state variable without APT present
- **z: Attack form** – previously detected class of APT effects on the state variable
- **p: Predict form** – APT effects on a state variable previously detected
- **o: unknown APT** affecting a given state variable

Note that **v**+**z** = **p**. If an unknown APT is similar to a previously seen APT, then **o**−**p**=**q** where if **q**=0 indicates the presence of APT and is referred to as a

Zero form, and **q**≠0 indicates that the APT is new and previously unknown but has now been detected.

The usefulness of prediction is in the application of previously developed mitigation methods for APT signatures of previously known APTs and the similarity of an unknown to a known APT. This is part of continuing and future work to define and evaluate.

Spicule Approach

Analytic vector mathematics can be used to redefine the mathematics in a spatial extent (Damballa, 2010). This type of visual rendering is not diagrams or pictures; it has an analytic algebra that can be utilized to analyse data as discussed in the previous section (**v, z, p, o, w, y**). This is the concept behind the development of a data representation of high order state variable vector data called Spicule (G. Vert, 2012). Spicule does its analysis based on concepts presented in the previous section based on modelling variables, referred to as state variables that describe some aspect of a system's or networks operation. It is possible to analyze up to tens of thousands of individual state variables and their changes to determine APT presence using the concepts previously presented. The analysis starts with population of state variable vectors around the radius of a Spicule in as small a degree increment as is required. Analysis for change (APT presence) is almost instantaneous using the fastest computational operation on a computer: integer addition and subtraction of vectors. Spicule's mathematical model and underpinning is based on a vector calculus. Its algebraic visual model can do the following:

- **Detect (c)** changes to a system instantly by only visualizing what has changed in the system. This facilitates human interpretation of the significance of the change and its potential threat. It also lends to automatic response and classification of malware activity.
- **Predict (p**) what a system will look like under attack.
- **Identify (z)** the essence of how an attack changes a system.
- Determine (**Zero** form to be discussed) if the states of a system or network have changed or not. This is a similarity operator in the algebra.

The Spicule while mathematically analytic, does consider the need to have analysts involved in the analysis of large state variable data sets in an easy to understand fashion. The interface is simple and intuitive for humans to interpret as shown in the next section. It lends nicely to interpretation of events in a system facilitating human ways of reasoning about and interpreting a possible APT attack (e.g., "most likely APT," "no APT," and so on). In essence the needle in the haystack can be found easily using the vector approach coupled with human interpretation of the vector states previously discussed.

Spicule Operation

While the goal of this effort is not to produce a prototypical visualization system for APTs, it is self-evident from previous work that the visualization is a useful tool in the interpretation of large amounts of complicated data. The Spicule is displayed as a sphere with two types of state variable vectors. There can be almost an infinite number of these vectors representing thousands of state variables for a given host or network of hosts. The two types of vectors are defined as:

- Fixed vectors (vector **G** in figure 1) that represent state variables ranging from 0 to ∞; for example, the number of users that are logged on the system.
- Tracking vectors (vector **B** in figure 1) that range in value from {0-100}%; for example, CPU usage.

Each vector is located at a degree location around the equator of the Spicule ball and represents a state variable that is being monitored for change. In a simple case, with tracking vectors ranging from 0 to 90 degrees located 360 degrees around the equator, and the tip of each tracking vector indicating a state the system is in or state variable value, it is possible to model 32,400 unique states at any given moment in time and instantly analyze change between Spicules from two moments in time to see if malware is active using the Zero form in figure 5. Subdivision of degree locations for the vectors around the equator leads mathematically to an almost infinite number of states that could be modelled.

A Zero form, shown in Figure 5 (at the bottom as a round featureless ball), results when a Spicule at time t1 is subtracted from a Spicule at time t0 and no change has occurred in state variables being modelled by the tracking and fixed vectors. A Zero form indicates that no malware is in operation and the state of the system has not changed. Research beyond the scope of this paper is defining methods to say vectors are similar even if they do not have the exact same value. This phenomenon is referred to as "jitter" and

deals with false positive and negative rates.

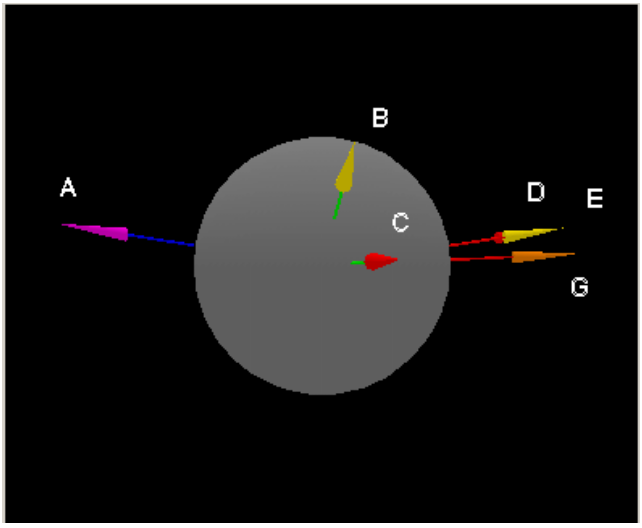


FIG. 1 SPICULE SHOWING SEVERAL PORTS ON A SYSTEM

The Spicule approach is to display system activity or state variables in the form of state variable vectors which are projected from the center of a sphere as in Figure 1. These vectors move or track as changes occur over time in a system. For example, a vector may represent CPU usage which can range from 0% to 100%. A CPU usage vector would normally start out at the equator to denote low CPU usage, but if the system found itself in the middle of a DOS (denial of service) attack, that same vector would be translated to pointing out of the northern pole to denote high CPU usage (near 100%).

For the initial development of a working Spicule prototype, we chose to test the concept by monitoring ports. While this prototypical test is not directly targeted at the realm of APTs specifically, it does serve to illustrate the early concept and so is included. Figure 1 is the vector display of a system monitoring a few state variables such as ports 22 (SSH, labelled A), 23 (Telnet, labelled B), 80 (HTTP, labelled C), 110 (POP3, labelled D), 137 (NetBIOS Name Service, labelled E), and 443 (HTTPS, labelled G).To test the prototype Backdoor SubSeven was used to simulate attack activity on specific ports. Backdoor SubSeven is a well-known Trojan. SubSeven works by opening an arbitrary port, specified by the attacker, but it most commonly attacks ports 1243, 6776, and 27374. Figure 2 shows the same system as before except that it is now under attack from SubSeven. The difference between these two Spicules is this new purple-tipped vector (labelled H) which has appeared suddenly with a great deal of traffic on an otherwise reserved port (1243).

Figure 1 is called a Normal form and Figure 2 is called a Change form, because something has changed in the vector states (the attack). The mathematics of calculating the Attack form and the relative reduction of data and interpretation of change is illustrated graphically in Figure 3. This shows the essence of the Backdoor SubSeven attack and its effects on the system.

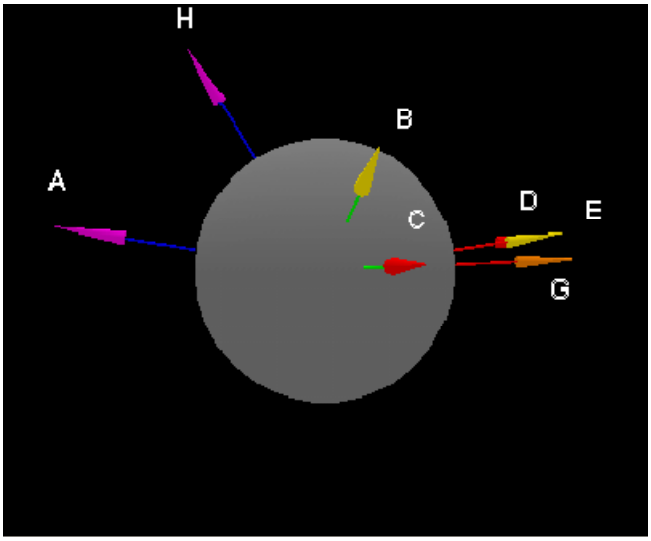


FIG. 2 SPICULE SHOWING A SYSTEM UNDER A SUBSEVEN ATTACK

Spicule Mathematical Properties and Analytic Vector Algebra

The Spicule model is comprised of six unique states: Normal form, Zero form, Change form, Attack form, Observe form, and Predict form. All of these are generated utilizing the previously discussed section on vector algebra and mathematics.

Definitions:

Normal form, N- a network systems state variables in normal ranges of operation

Change form, C - a systems state variables that have been changed

Attack form, A - the representation of only the state variables modelling malware effects on the system

Predict form, P - a representation of what a Normal form would look like given a specific malware operation

Zero form, Z - a featureless Spicule indicating that two forms used to create it are exactly the same

Observe form, O - the result of an algebraic operation, can be the same as a Change form or a Zero form

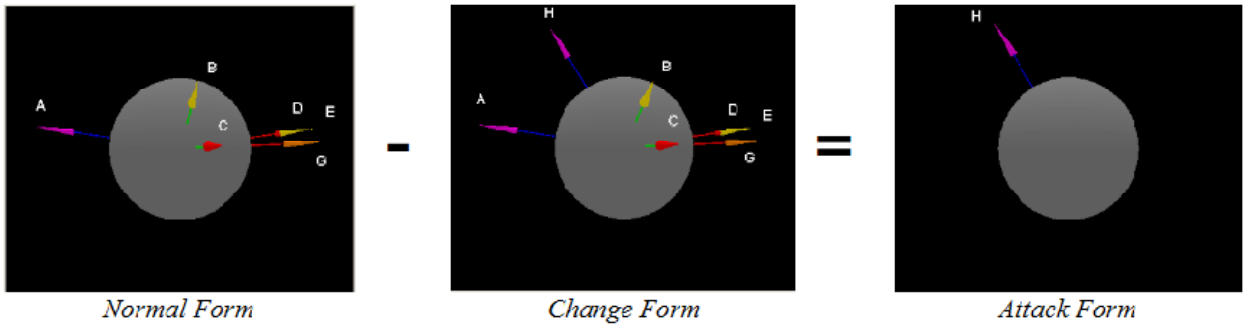


FIG. 3 CALCULATING THE ATTACK FORM

It is important to note the type of the form depends on the types of the forms used to calculate it and the algebraic operation used in the calculation. The algebra and calculations are presented in the following sections.

The **Normal (N) form** is the state in which the system is operating normally and not under attack. Opposite to this is the **Change form (C),** which is a representation of a system under attack. The **Attack form (A**) is a signature (or isolated) view of an attack in progress that is occurring inside the Change form. The Attack forms can be stored in a database for later reference, in which case they become Predict forms which are predictions of potentially future attacks. The Observe form is a state which may or may not be an attack signature. Through the vector calculus, an Observe form can be compared to a Predict form. Each one of these forms has a unique appearance and mathematical signature.

Most operations to produce the above forms are accomplished by adding two forms (their state variable vectors) together or subtracting one from another. The algebra is performed by iterating through the vectors of each Spicule and performing individual vector operations depending on the algebraic function being calculated. For example, in order to isolate an attack and produce an Attack form, simply subtract the Change form from the Normal form as follows:

$$\mathbf{A=N-C} \quad (1)$$

where A is the Attack form of Spicule, N is the Normal form, and C is the Change form as shown in figure 3. The algorithm for this above process is:

FOR EACH v *on the Spicule:*
$A_v = N_v - C_v$
END FOR.

In figure 3, one can see the essence of the attack's visual characteristics in the Attack form (A). Such forms can be potentially stored into a database as the Attack form of SubSeven or the family of malware that operates similar to SubSeven. Once they become stored and classified, they become a Predict form used in the future to identify the same or similar classes of attacks. Attack forms can be created from pre-classification of attack families for the major families of malware or APT. They are stored and used for identification of future attacks as Predict forms. They are subtracted from a Change form to classify an attack if a Zero form results from the algebra. The Attack form of Spicule is a classification of a type or family of attacks based on how they change the system state variables over time.

An Observe form may or may not represent an Attack form. It is generated by subtracting Spicules at different points in time to see if any change vectors appear. It can then be subtracted with a Predict form stored in a database to classify the family of attack that is occurring on the system.

A Change form is always Spicule at time t_1 that a Normal form (at time t_0) is subtracted from to calculate the Observe form. To summarize, one can detect an attack by using the following equation:

$$O=N-C \quad (2)$$

where O is the Observe form, N is the Normal form, and C is the Change form. The major difference between equations (1) and (2) is that the latter is used to create an observe form, which is a possible Attack form, whereas the former is used when creating an Attack form only. The reasoning behind this is that (1) will be used to create a library of all attacks ever witnessed, and the result of (2) will be used to detect an attack underway against attacks stored in our library.

The Observe form is potentially what an attack would look like while it is underway. It is compared against

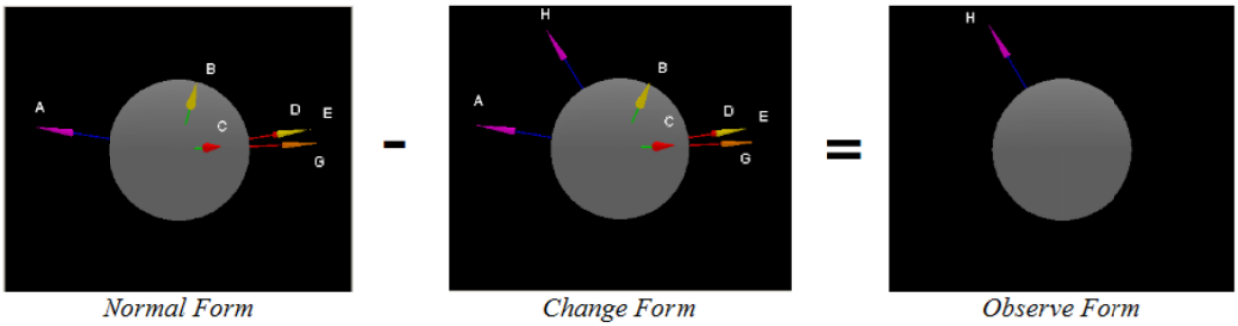


FIG. 4 CALCULATING THE OBSERVE FORM

the stored known Attack forms to identify an attack. The method of performing this comparison is an algebraic subtraction as shown in the following:

$$Z=A-O \quad (3)$$

where Z is the Zero form, A is the Attack form, and O is the observe form. Essentially the Zero form is a Similarity operation. Figure 5 shows the actual Spicules applied to the Observe from figure 4.

A Predict form is meant to determine what a system might look like if a given attack from a family of malware is present on the system. It is way to watch for such an event if it occurs. The Predict form is created by the additive property of the algebra. It is calculated as the following:

$$P=N+A \quad (4)$$

This produces what we expect an attack or APT activity to look like if it occurs. The subtraction operation, the similarity operation creating the Zero form then identifies and confirms that the APT or attack malware is present via:

$$Z=P-C \quad (5)$$

If a Zero form exists then the attack has been identified, classified, and can be responded to as shown in figure 5. It is important to note that a Zero form occur with subtraction of one set of state variable vectors from another when they exactly match or jitter control has been applied. Figure 5, shows the Zero form on the right which means that the Predict or Attack forms match the current changes in the system, the Observe form. Note that similarity results in a featureless Spicule hence the name Zero form. This drastically simplifies and speeds the process of recognition by analyst. Mathematically interpretation can be also automated because the result of subtraction produces a Z = (0, 0, 0). The potential gain in identification time has the possibility to extend Spicule methodology to real-time visual and/or automated detection.

FAST-VM, Finite Angular State Transition Velocity Machine Model Concept

The next part of our research extends Spicules vector algebra into a state machine that can model and analyze thousands of state variables over time and space. Time and spatial would be the type of modelling FAST-VM would perform for a large complex network operating over a period of time. This is appropriate because the challenge in APT detection is the sophistication of the threats operation and the stealth over time. Watching for a state variable change at some point in time is not sufficient. It is conjectured that there is a need to watch a very large number of state variables and then distill the essence of the APT's activities using the sort of high order data set reduction previously shown in the Spicule approach. If an attacker has an effect on a machine, the attacker will also change its state in subtle ways. These ways are not easily predictable unless the Spicule model is adapted to look at the velocity (rate) of state changes over time. This is the essence of the FAST-VM concept.

FAST-VM builds on Spicule algebraic analysis capabilities. It models a signature of Spicules as they transition from one state form for a portion of an attack to another state form as shown in Figure 6. The FAST-VM model transitions to new states which are measured by their state transition velocities. Velocity still a part of future research but informally is defined to be the rate of change in thousands of vectors over time to a new Spicule state. FAST-VM signatures or branches can generate in any direction in 3D space. Considering that a branch of a FAST-VM transition as shown in figure 6 could be located at every single degree location from an initial Spicule state, it is possible to model potentially huge amounts of data and variations of malware or APT operation. When an APT starts to generate a FAST-VM signature branch it can be compared to existing branches stored in a database to determine if it is similar to previous

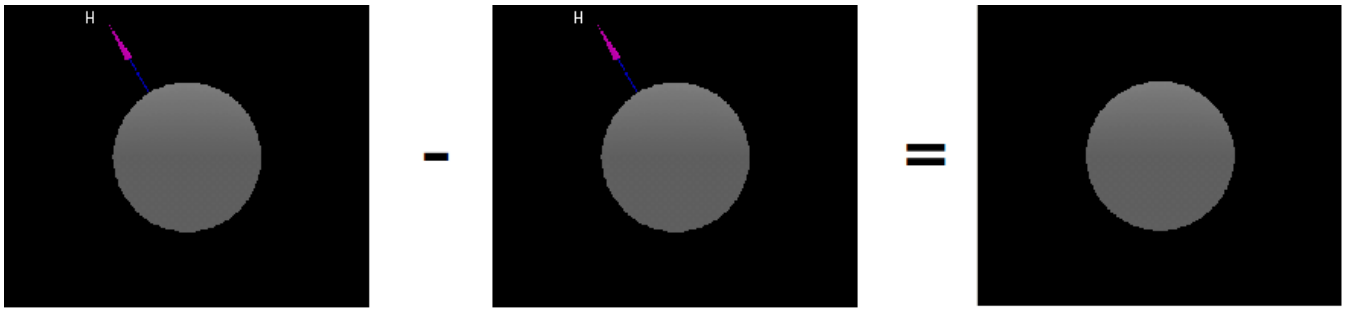


FIG. 5 CALCULATING THE ZERO FORM

patterns of activity. This comparison can allow a system to classify known and similar types attacks in a computationally efficient because it is based on integer vector data.

The FAST-VM models high order data spaces of state variables as would be found on a large network that could be affected by an APT and reduces them to the essence of an APT's effect on state variables as its operates over time. It has the ability to model thousands of state variable vectors and then perform the following key functions from the Spicule model:

- **Find** the needles in the haystack representing the APTs' effects on the network through detection of changes in high order state variable data sets, (3, 2) the **Observe** form.
- **Detec**t previously unknown APTs (1), the **Attack** form
- **Classify** state activity changes of unknown APTs as similar to known APTs allowing mitigation methods for a known APT to be applied to an unknown APT once the FAST-VM has detected its presence, (5) the **Zero** form.
- **Predic**t what a category of an APT's attack on a system might look like so that it can be monitored for presence, (4) the **Predict** Form.

FAST-VM Operation

The FAST-VM model consists of a series of Spicules as they transition over time that can be combined into N-dimension state transitions or an elaborate graph. Each transition has a velocity which is still being defined. The velocity is defined by the rate of change in Spicule Change forms over time and possibly an attribute of probabilistic confidence that denotes the transitioned to state as a recognized state. In each state of the graph created by transitions to a new state, the Spicule algebra is applied, evaluating Change forms, and applying Predict form analysis. If a series of Predict forms match what is observed, then the APT can be classified and identified. The model of FAST-VM's operation is shown in Figure 6.In this example, one can see the following characteristics:

- Spicules representing state changes in state variable at various points in time;
- a velocity equation |h| (magnitude of the transition) that describes the transition speed from t_0 to t_1 ;
- a cumulative Attack form describing the attack signature for APT summed over time at t_3 ;
- a Bayesian probability P based on confidence that the attack signatures are known to be part of the transition attack profile for malware, where M_x is malware x and P_n is the probability of having a known attack Spicule form at T_n. P is thought to deal with the issue of jitter, in that for a given malware family, Spicule Attack forms at any given point in time should be similar but may not be exactly the same.

Figure 6 shows the operational aspects of FAST-VM with the Spicule algebra calculating change forms

where:

T_n : time slice

|h| : velocity as a function of magnitude of transition to a new Change form

P (M_x| P_{to}, P_{t1}, P_{t2}): probability malware APTx given probability time slices model components of unknown system operation, or known/similar APTx state transitions.

The above diagram shows a FAST-VM creating a *signature trail* as an APT is being analyzed at each step in the process. This trail considers the rate of change to a new state over time and the high order of state variables that can evaluated and reduced to the essence of the attack (Attack form or Change Forms) as each transition. Of note, only Change Forms are being

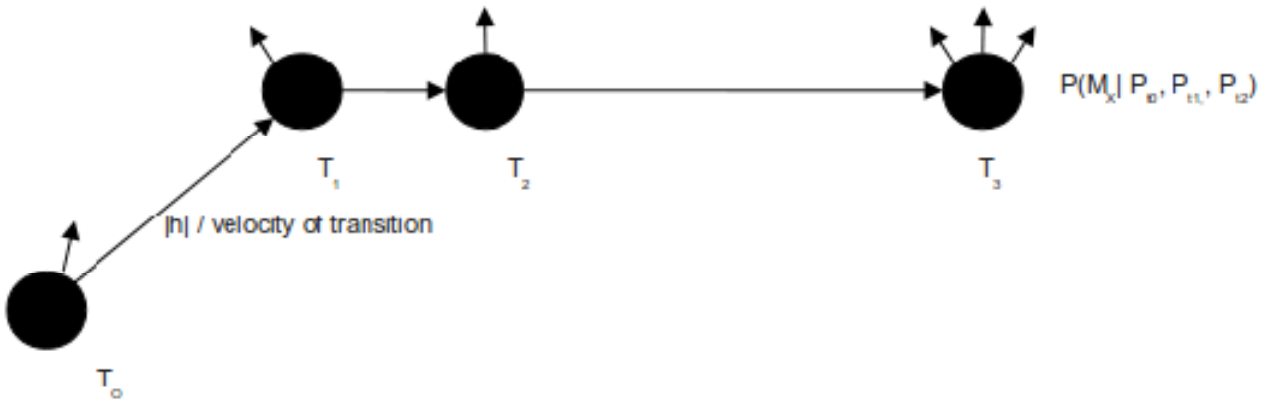


FIG. 6 EXAMPLE FAST-VM TRANSITIONING AND SPICULES ANALYSIS FOR APT OVER TIME IN A NETWORK OR ON A HOST

modelled in figure 6. In reality each Spicule could have thousands of state variables that have not changed and thus are eliminated in the vector algebraic Spicule analysis.

Performance Testing

Finally, one of the key features of FAST-VM is its ability to model and analyze large amounts of data. This should be done with as low computational overhead as possible. The vector operations in the algebra were modelled in a simulation using Visual Studio 2008 on an Intel laptop with fours cores running Windows 7. The results are shown in the following tables

TABLE I DIFFERENCING CALCULATIONS ("-")

Number of Vectors Evaluated (Integer)	Time to Execute (ms)
1000	0
1,000,000	1
1,000,000	3
10,000,000	24-58 avg around 42
100,000,000	286

TABLE II ADDITIVE CALCULATIONS ("+")

Number of Vectors Evaluated (Integer)	Time to Execute (ms)
1,000	0
100,000	0
1,000,000	3
10,000,000	25
100,000,000	289

The resultant calculations suggest that FAST-VM could analyze and model large amounts of data across a complicated and active network with relatively low computational overhead. This will be the subject of future evaluation.

Conclusions and Future Research

Detection of sophisticated slow moving threats such as APT requires the ability to model large amounts of data and reduce it rapidly to the essence of what has changed in system state over time. Analysis then has to have the capability to create signatures that can be stored for normal state transition and compared against to predict future unknown but similar types of attacks. FAST-VM and its vector based analytic algebra now provide a potential powerful model for detection of advanced threats. As shown in the preliminary performance results FAST-VM can computationally efficiently analyze large amounts of data and can scale to network detection of hundreds of thousands of network states. In this theoretical model APT was chosen as a vehicle to demonstrate the FAST-VM model and how its theoretical underpinnings operate. There are many research areas still to be investigated to develop the model further.

The first of these areas involves research that centers on the isolation of state changes indicative of APT presence. There is also a degree of classification work that needs to be completed that attempts to define and quantify broad categories of APTs that share predictive characteristics. Another future research task is involved in the refinement of techniques that predict the expected state of a system; that is, there is no attack and things are operating normally. As this is intended to be prototypical and demonstrative, it is expected that, to a degree, our prediction will be tailored to a demonstration prototype system initially. Included in this is the identification of a large collection of state variables describing host and network operation from ongoing and past research.

Research needs to be further conducted to evaluate the adaptive jitter methods to deal with vectors that do not have exactly the same locations but are essentially the same. This phenomenon is referred to as jitter. A theoretical model is in development to deal with reduction of false positives and false

Finally, hyper scalability needs to be further modelled

and developed for networks of FAST-VMs. It is thought that cooperative colonies of FAST-VM's around a large network might be a promising approach to autonomous threat detection and perhaps mitigation.

REFERENCES

Damballa, "What's an APT? A Brief Definition". January 20, 2010.

Dell SecureWorks, "Anatomy of an Advanced Persistent Threat (APT)". Retrieved 2012-05-21

Erbacher, R. Walker, K. Frincke, D. Intrusion and Misuse Detection in Large-Scale Systems. IEEE Computer Graphics and Applications. Feb, 2002.

G. Vert, Anitha Chennamaneni, S.S. Iyengar. A Theoretical Model for Probabilistically Based Detection and Mitigation of Malware Using Self Organizing Taxonomies, SAM 12, Las Vegas, July 2012.

G. Vert, D. Frincke. Towards a Mathematical Model for Intrusions. NISS Conference. 1996.

G. Vert, Frincke, D.A., and McConnell, J., "A Visual Mathematical Model for Intrusion Detection," Proceedings of the 21st NISSC Conference, Crystal City, Virginia, 1998.

G. Vert, S. Sitharama Iyengar, Vir V. Phoha: Defining A New Type of Global Information Architecture for Contextual Information Processing. IKE 2009: 214-219.

Han, Gensuo and Kagawa, Koji. "Towards a Web-based program visualization system using Web3D." Paper presented at the meeting of the ITHET, 2012.

James, J. Bricken, W.A boundary notation for visual mathematics. Visual Languages, 1992. Proceedings. Sep, 1992.

Kareev, Y., Fiedler, K., & Avrahami, J. (2009). Base Rates, Contingencies, and Prediction Behavior. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35 (2), 371-380.

Michael K. Daly, "Advanced Persistent Threat (or Informationized Force Operations)". Usenix. November 4, 2009.

Mohammad Sazzadul Hoque, Md. Abdul Mukit, Md. Abu Naser Bikas.AN IMPLEMENTATION OF INTRUSION DETECTION SYSTEM USING GENETIC ALGORITHM. International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.2, March 2012.

Neil MacDonald. The Future of Information Security Is Context Aware and Adaptive. Gartner Research. May 15, 2010.

S. Eick and G. Wills. Navigating Large Networks with Hierarchies, In Proceedings Visualization Conference '93, pp. 204-210, San Jose, Calif., October 1993.

Scarfone, Karen; Mell, Peter. "Guide to Intrusion Detection and Prevention Systems (IDPS)". Computer Security Resource Center (National Institute of Standards and Technology) (800–94), 2012.

Shuo-Liangxun Zhao-Jinhui, Wang-Xuehui. An Adaptive Invasion Detection Based on the Variable Fuzzy Set. 2011 International Conference on Network Computing and Information Security (NCIS). May 2011.

Zulaiha Ali Othman, Azuraliza Abu Baker, Intesar Etubal. Improving Signature Detection Classification Model Using Features Selection based on Customized Features. Dec. 2010.

Gregory Vert earned a PhD from the University of Idaho. He holds a CISSP and CCENT certification and is the founder and Coordinator of the Security Education Research and Training (SERT) lab at Texas A&M Central Texas. He completed his post-doc at LSU in the Computer Science department in 2009 and was appointed to the Center for Secure Cyberspace.

He has authored 70+ papers with most of them in the area of security. He also defined and published a book recently on the topic of Contextual Processing and contextually based security.

Dr. Vert is a member of ACM and IEEE and was the founder of these organizations at Texas A&M.

Bilal Gonen was born in Istanbul, Turkey. He received a B.E. in Computer Engineering in Turkey in 2002. He received his M.S. in Computer Science from the University of Georgia, U.S.A. in 2006. He received his Ph.D. in Computer Science from University of Nevada, Reno, in Nevada, U.S.A. in 2011.

He is currently an Assistant Professor in Computer Science department at the University of West Florida, in Florida, U.S.A.After completing his Ph.D. he worked as a Term Assistant Professor at University of Alaska, Anchorage (Alaska, U.S.A.). His research interests include Network

Science, Social Network Analysis, Semantic Web, Computer Networks, and Bioinformatics.

Dr. Gonen is a member of ACM, IEEE and Society for Science & the Public (SSP).

Jayson G. Brown was born in Abilene, Texas USA. He received his Master of Science in 2013 from Texas A&M University – Central Texas in the major of Information Systems as a Distinguished Student. His previous undergraduate background involves Computer Information Systems in Software Engineering with Summa Cum Laude honors.

He has conference proceedings relating to knowledge transfer and software methodologies. He is in the midst of finding a PhD program at the current time, but is maintaining employment as a Graduate Assistant with the CIS department. His personal research interests include software evolution, biometrics, AI, machine learning or other similar fields.

Jayson is a member of ACM, AMCIS and various honor societies including Delta Mu Delta.

doi: 10.14355/ijcsa.2014.0302.02

A New Approach for Subspace Clustering of High Dimensional Data

[1]Mrs. M. Suguna, [2]Dr. S. Palaniammal

[1]Assistant Professor, Department of Computer Science, Dr. GRD College of Science, Coimbatore.

[2]Professor & Head, Department of Science & Humanities, Sri Krishna College of Technology, Coimbatore.

[1]Suguna_a@yahoo.com

Received 8 January 2014; Accepted 9 February 2014; Published 15 May 2014

Abstract

Clustering high dimensional data is an emerging research area. The similarity criterion used by the traditional clustering algorithms is inadequate in high dimensional space. Also some of the dimensions are likely to be irrelevant thus hiding a possible clustering. Subspace clustering is an extension of traditional clustering that attempts to find clusters in different subspaces within a dataset. This paper proposes an idea by giving weight to every node of a cluster in a subspace. The cluster with greatest weight value will have more number of nodes when compared to all other clusters. This method of assigning weight can be done in two ways such as top down and bottom up. A threshold value is fixed and clusters with value greater than threshold only will be taken into consideration. The discovery of clusters in selected subspaces will be made easily with the process of assigning weight to nodes. This method will surely result in reduction of search space.

Keywords

Clustering; Curse of Dimensionality; Subspace Clustering; High Dimensional Data

Introduction

Clustering is an established technique in knowledge discovery process. Clustering aims at dividing data into groups such that objects in one group are similar to each other with respect to similarity measure and objects in the other groups are dissimilar. Traditional clustering techniques may suffer from the problem of discovering meaningful clusters due to the curse of dimensionality. Curse of dimensionality [P. Indyk and R. Motwani, 1998], refers to the phenomenon that as the increase of the dimension cardinality, the distance of a given object *a* to its nearest point will be close to the distance of *a* to its farthest point. A feature transformation and feature selection method addresses this problem by reducing the dimension. Feature transformation techniques summarize the data by taking linear transformation of attributes. Transformations preserve the original distance between objects. However the resulting clusters are hard to interpret. Feature selection methods work at the principle of reducing dimension. They determine relevant attributes such that only a particular subspace is selected. However, clusters may be embedded in varying subspaces. Due to this fact a significant amount of research has been done on subspace clustering, which aims at discovering clusters embedded in any subspace of the original space [Jiawei Han and Michaline kamber, 2010].

Some of the significant areas of subspace clustering are information integration systems, text-mining and bioinformatics. Subspace clustering has been applied in various applications like gene expression analysis, E-commerce, DNA microarray analysis and so on. Most of the previous subspace clustering works [E. Muller, S. Gunnemann, I. Assent, and T. Seidel, 2009], [H. –P. Kriegel, P. Kroger, and A. Zimek, 2009] by regarding clusters as regions of high density regions than their surroundings. This paper works on the basis of assigning weight to every node of a cluster in a subspace. The cluster with greater weight value will have more numbers of nodes when compared to other clusters.

Related Work

Clique

CLIQUE [R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, 1998] starts with the definition of a unit – elementary rectangular cell in a subspace. Only units

whose densities exceed a threshold are retained. A bottom-up approach of finding such units is applied. First, 1-dimensional units are found by dividing intervals in û equal-width bins (a grid). Both parameters and û are the algorithm's inputs. The recursive step from q-1-dimensional units to q-dimensional units involves self-join of q-1 units having first common q-2 dimensions (Apriori-reasoning). All the subspaces are sorted by their coverage and lesser-covered subspaces are pruned. A cut point is selected based on MDL principle. A cluster is defined as a maximal set of connected dense units. It is represented by a DNF expression that is associated with a finite set of maximal segments (called regions) whose union is equal to a cluster. Effectively, CLIQUE results in attribute selection (it selects several subspaces) and produces a view of data from different perspectives. The result is a series of cluster systems in different subspaces. This versatility goes more in vein with data description rather than with data partitioning. There will be overlapping of clusters. If q is the highest subspace dimension selected, the complexity of dense units generations is O ($const^q$ + Nq). Identification of clusters is a quadratic task in terms of units.

Enclus

The algorithm ENCLUS (ENtropy-based CLUStering) [C.H. Cheng, A.W. Fu, and Y. Zhang, 1999] follows in the footsteps of CLIQUE, but uses a different criterion for subspace selection. The criterion is derived from entropy related considerations: the subspace spanned by attributes A1,..., Aq with entropy H (A1,..., Aq) smaller than a threshold τ is considered good for clustering. Any subspace of a good subspace is also good, since

$H(A_1,...,Aq_{-1}) \leq H(A_1,..., A_{q-1}) + H(A_q \mid A_1,..., A_{q-1}) = H(A_1,..., A_q) < \omega$.

Low entropy subspace corresponds to a skewed distribution of unit densities. The computational costs of ENCLUS are high.

Mafia

The algorithm MAFIA (Merging of Adaptive Finite Intervals) [H.S. Nagesh, S. Goil, and A. Choudhary, 2001] significantly modifies CLIQUE. It is a technique for adaptive computation of finite intervals. It starts with one data pass to construct histograms. Adjacent bins with similar histograms are combined to form larger bins. Cells with low density will be eliminated thereby reducing the computation. The algorithm uses a parameter *Į*, called cluster dominance factor, to select bins that are *Į*-times more densely populated relative to their volume than on average. These are q=1 candidate dense units (CDUs). Then MAFIA proceeds recursively to higher dimensions (every time a data scan is involved). The difference between MAFIA and CLIQUE is that to construct a new q- CDU, MAFIA tries two q-1-CDUs as soon as they share any (not only first dimensions) q-2-face. This creates an order of magnitude more candidates. Adjacent CDUs are merged into clusters and those clusters that are proper subsets of the higher dimension clusters are eliminated. The parameter *Į* with default value of 1.5 works fine. Reporting for a range of *Į* in a single run is supported. If q is a highest dimensionality of CDU, the complexity is **O (constq + qN).**

Predecon

The algorithm PreDeCon (PREference weighted DEnsity CONnected clustering) performs clustering with a single scan over the database. Each object is marked as unclassified at the beginning. During successive runs, all the objects are assigned either a cluster identifier or marked as noise. PreDeCon [Christian Bohm, Karin Kailing, Hans-Peter Kriegel, Peer Kroger, 2004] checks whether the unclassified point is a preference weighted core point. If so, the algorithm expands the cluster belonging to this point. Otherwise it is marked as a noise. The complexity of the algorithm based on the sequential scan of the data set is O ($d.n^2$).

Subclu

The algorithm SUBCLU (density-connected SUBspace CLUstering) [K. Kailing, H.-P.Kriegel, and P. Kroger, 2004] is able to detect arbitrarily shaped and positioned clusters in subspaces. It is based on the bottom-up greedy approach. It starts with the generation of all 1-dimensional clusters by applying DBSCAN to each 1-dimensional subspace. For each k-dimensional subspace S, the algorithm search for all other k-dimensional subspace T having (k-1) attributes in common. It then joins them to generate (k+1) dimensional candidate subspaces. The algorithm prunes 1-dimensional subspaces thereby reducing the number of candidate subspaces. In this way k+1 clusters are formed.

Edsc

The algorithm EDSC (Efficient Density-based Subspace Clustering) [Ira Assent, Ralph Kriegel, Emmaneul

Muller, and Thomas Seidl, 2008] uses a multistep filter-and-refine approach. The first filter step determines the hypercube that surround each possible subspace cluster. Then candidates with higher dimensional projection are processed. The second filter step is the density filter which performs lossless pruning with respect to the unbiased density. The last step is the refinement step which undergoes determination of subspaces with minimum density criterion.

Dencos

The algorithm DENCOS (DENsity Conscious Subspace clustering for High-dimensional data) [Yi-Hong Chu, Jen-Wei Huang, Kun-Ta Chuang, De-Nian Yang, and Ming-Syan Chen, 2010] implements divide-and-conquer scheme to discover clusters that satisfy different density thresholds in different subspace cardinalities. Identification of dense units is similar to frequent itemset mining. Dataset is first preprocessed by transforming each *d*-dimensional data point into *d* one-dimensional units. DFP-tree is constructed on the transformed data set that stores complete information of the dense units. The computation of lower bound and upper bound of the unit counts from the constructed DFP tree helps in speeding up dense units discovery. The quality of clustering results is evaluated in terms of density ratio.

Proclus

The algorithm PROCLUS [Charu C. Aggarwal, Cecilia Procopiuc, Joel L Wolf, Philips S. Yu, Jong Soo Park, 1999] (PROjected CLUstering) associates with a subset C a low-dimensional subspace such that the projection of C into the subspace is a tight cluster. The subset – subspace pair when exists constitutes a projected cluster. The number k of clusters and the average subspace dimension l are user inputs. The iterative phase of the algorithm deals with finding k good medoids, each associated with its subspace. A sample of data is used in a greedy hill-climbing technique. Manhattan distance divided by the subspace dimension is a useful normalized metric for searching among different dimensions. An additional data pass follows after iterative stage is finished to refine clusters including subspaces associated with the medoids.

Orclus

The algorithm ORCLUS (ORiented projected CLUSter generation) [Charu C. Aggarwal, Philips S. Yu, 2000] uses a similar approach of projected clustering, but employs non-axes parallel subspaces of high dimensional space. In fact, both developments address a more generic issue: even in a low dimensional space, different portions of data could exhibit clustering tendency in different subspaces (consider several non-parallel non-intersecting cylinders in 3D space). If this is the case, any attribute selection is doomed. ORCLUS has a k-means-like transparent model that defines clusters as sets of points (partition) that have low sum-of-squares of errors (energy) in a certain subspace. More specifically, for $x \in C$, and directions $E = \{e1,...,ej\}$ (specific to C), the projection is defined as {x e1,...,x ej}.

The projection only decreases energy. SVD diagonalization can be used to find directions (eigenvectors) corresponding to the lowest l eigenvalues of the covariance matrix. In reality, the algorithm results in X partitioning (the outliers excluded) into k clusters Cj together with their subspace directions Ej. The algorithm builds more than k clusters with larger than l-dimensional E gradually fitting the optimal subspace and requested k. Though suggestion of picking a good parameter l is provided, uniform l is a certain liability.

Our Contributions

The proposed procedure attempts to overcome the curse of dimensionality problem. The procedure starts with the determination of maximal dense regions over the datasets in the subspaces. Maximal dense units are discovered by considering the item set as a graph D with nodes N and E edges. The maximum density present inside the graph is determined and that will be treated as the subgraph of the graph. Weights such as a1, a2, a3, etc are assigned to each node. The problem is to identify the node with maximum intensity. Those nodes with the highest value of weight are treated maximal dense nodes. This process is carried out after assigning weight to all the nodes. Bottom-up approach is used to find the subgraph with maximum intensity. Top-down method is followed to assign weight to the nodes. Binary partitioning method is applied to traverse all the nodes which reduce the search space about 50% for each iteration.

Proposed Algorithm

Consider D be the graph with N nodes and E edges. Let p be the dimension, following is the algorithm to find dense sub graph:

1. For each dimension *p* in D

2. Partition *d* into equal intervals [like for doing binary

search]

3. Fix a minimum and maximum threshold o (by applying APRIORI property).

4. Find the dense units.

5. S*et threshold = true*

6. While(true)

7. For each combination of *m* dimensions *p1, p2,……..,pm.* for each intersection *j* of dense units along *m* dimensions.

9. If *j* is dense

10. Identify *j* as a dense unit.

11. If no more units are identified as dense, break the loop.

After computing the cluster, we have to set the threshold to find the number of cluster falls under the particular threshold. This cluster will contain high dimensional data.

Clustering is a process of grouping number of similar things according to characteristics or behaviors. This can be used to cluster high dimensional data present in large tables.

We have to formulate another procedure to reduce the space used by the data inside each clustering.

Principle Component Analysis (PCA) is also a good idea to reduce curse of dimensionality. This can be applied after our procedure of giving weight to the item sets and finding all the sub clusters to get better solution. It is the most traditional tool used for dimension reduction. PCA projects data on a lower-dimensional space, choosing axes keeping the maximum of the data initial variance. PCA is a linear tool. There are several ways to design nonlinear projection methods. The first one consists in using PCA, but locally in restricted parts of the space. Joining local linear models leads to a global nonlinear one.

Principal component analysis (PCA) is a popular technique for dimensionality reduction which, in turn, is a necessary step in classification. It constructs a representation of the data with a set of orthogonal basis vectors that are the eigenvectors of the covariance matrix generated from the data, which can also be derived from singular value decomposition. By projecting the data onto the dominant eigenvectors, the dimension of the original dataset can be reduced with little loss of information.

In PCA-relevant literature, the covariance matrices are constructed using the eigenvalue /eigenvector approach. To perform efficient computation PCA is being used with singular value decomposition of the data matrix. The relationship between the eigen-decomposition of the covariance matrix and the SVD of the data matrix, is presented below. In this paper, eigen-decomposition of the covariance matrix and SVD of the data matrix are used interchangeably.

Let X be the data repository with m records of dimension d (m £d). Assume the dataset is mean-centered by making $E[X] = 0$. A modern PCA method is based on finding the singular values and ortho-normal singular vectors of the X matrix.

Using covariance matrix to calculate the eigenvectors, let $C = E[X^T, X]$ represent the covariance matrix of X. Then the right singular vectors contained in V are the same as those normalized eigenvectors of the covariance matrix C. In addition, if the nonzero eigenvalues of C are arranged in a descending order, then the kth singular value of X is equal to the square root of the kth eigenvalue of C.

With the help of PCA and SVM we can perform the followings:

(1) The classification boundary functions of SVMs maximize the margin, which equivalently optimize the clusters present inside the graph.

(2) SVMs handle a nonlinear classification efficiently using the kernel trick that implicitly transforms the input space into another higher dimensional feature space.

(3) It can also be used to sort out the outliers present inside the graph. The outliers are the unwanted data which will crease the space of the graph without providing any use. To reduce outliers we can form rules such that those nodes in the cluster not satisfying the rule will be dropped from the same.

Conclusion

In this paper, a new algorithm has been proposed for reducing subspaces with a threshold value. It also provides an idea of different mining algorithms and their procedures and how this new algorithm differs from the existing algorithms. It performs better than other algorithms on different criteria basis. The future work will deal with different data sets and comparison of results to estimate the efficiency of this new algorithm.

REFERENCES

C.H. Cheng, A.W. Fu, and Y. Zhang, *"Entropy-Based Subspace Clustering for Mining Numerical Data,"* Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 1999.

Charu C. Aggarwal, Cecilia Procopiuc, Joel L Wolf, Philips S. Yu, Jong Soo Park, "Fast Algorithms for Projected Clustering ", *SIGMOD*, 1999.

Christian Bohm, Karin Kailing, Hans-Peter Kriegel, Peer Kroger, "Density connected clustering with local subspace preferences", *ICDM*, 2004.

Charu C.Aggarwal, Philips S. Yu, "Finding generalized projected clusters in high-dimensional spaces" 2000.

D. W. Scott, J. R. Thompson, Probability density estimation in higher dimensions. In: J.E. Gentle (ed.), Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface, Elsevier Science Publishers, 1983, pp.173-179

E.Muller, S.Gunnemann, I.Assent, and T.Seidel, "Evaluating clustering in subspace projections of high dimensional data", in VLDB, 2009, pp. 1270 – 1281.

H.–P.Kriegel, P.Kroger, and A.Zimek, "Clustering High Dimensional Data: A survey on subspace clustering, pattern-based clustering and correlation clustering", *TKDD*, vol. 3, no. 1, pp. 1-58, 2009.

H.S. Nagesh, S. Goil, and A. Choudhary, "*Adaptive Grids for Clustering Massive Data Sets,"* Proc. First SIAM Int'l Conf. Data Mining (SDM), 2001.

Ira Assent, Ralph Kriegel, Emmaneul Muller, and Thomas Seidl, "EDSC: Efficient Density-based Subspace Clustering", *CIKM*, pages 26-30, 2008.

Jiawei Han and Michaline kamber, *"Data Mining Concepts and Techniques"*, Morgan Kaufmann Publishers, 2nd Edition.

K. Kailing, H.-P. Kriegel, and P. Kroger, *"Density-Connected Subspace Clustering for High-Dimensional Data,"* Proc. Fourth SIAM Int'l Conf. Data Mining (SDM), 2004.

K. Kailing, H.-P.Kriegel, and P. Kroger. Density-connected subspace clustering for high-dimensional data. In *SDM*, pages 246-257, 2004.

K. Sequeira and M. Zaki. SCHISM: A New Approach for interesting subspace mining. In *ICDM* pages 186-193, 2004.

P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In Proc. 30th Symp. on Theory of Computing, pages 604–613. ACM, 1998.

R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, *"Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications,"* Proc. ACM SIGMOD Int'l Conf. Management of Data, 1998.

Yi-Hong Chu, Jen-Wei Huang, †Kun-Ta Chuang, and Ming-Syan Chen *"On Subspace Clustering with Density Consciousness"*.

Yi-Hong Chu, Jen-Wei Huang, Kun-Ta Chuang, De-Nian Yang, Member, IEEE, and Ming-Syan Chen, Fellow, IEEE, *"Density Conscious Subspace Clustering of High Dimensional Data"*, IEEE Transactions on knowledge and data Engineering, 2010.

International Journal of Computer Science and Application (IJCSA) Volume 3 Issue 2, May 2014 www.ij-csa.org
doi: 10.14355/ijcsa.2014.0302.03

Improving User Adherence to a Reminder: a Model and Methodology

Lei Tang*[1], Zongtao Duan[2], Hanbo Wang[3], Zhiwen Yu[4]

[1,2]School of information engineering, Chang'an University, Xi'an, China.

[3]Shaanxi regional electric power goup co., LTD, Xi'an, China

[4]School of computer science, Northwestern Polytechnical University, Xi'an, China

*[1]tanglei24@chd.edu.cn; [2]ztduan@chd.edu.cn; [3]wanghb@gmail.com; [4]zhiweny@nwpu.edu.cn

Received 25 October 2013; Accepted 19 December 2013; Published 15 May 2014

Abstract

Various reminder systems have emerged to promote users' health. However, it is hard to say such systems by which person is passively following the instructions can truly meet the goal. This means researchers are under pressure to find the ways of persuading a user to accept the instructions. Seeking the best time to provide the prompts is important to improve user adherence. A methodology is proposed by a following idea where the full benefit of the acceptable reminder will be achieved only if the user is open-minded to an interruption caused by it reasonably and closely. The time for prompting is thus evaluated before a decision on when a user is most receptive. We describe our experience in delivering the attention-sensitive reminders for elder adults to take their medication. Experimental results show that the approach can improve user adherence in an efficient and flexible way.

Keywords

Attention; Interruption; User Adherence; User's Activity; Reminder

Introduction

Pople begin to enjoy many of those pleasures benefiting from pervasive computing. Indeed, Weiser's vision [Weiser] will become a reality with various novel systems emerged as communal supports for human being. These systems have made different efforts to advocate people of enjoying the advantages of life by using the pervasive services provided at an appropriate time. However, they haven't yet realized the fact that such attractive services (e.g. assistance) have turned out to be the interruptions upon the users [Rivera-RodriguezKarsh]. These interruptions such as alarms, message arriving, equipment failures and other person's requests can distract user's attention, and bring a shift in attention. The goal-driven interruptions are essential to the task process and provide necessary information (e.g. an elder's monitor alarming due to abnormal vital signs). However, some literature [Drews] has also stated that those occur when users are expected to perform tasks requiring their undivided attention, that often disrupts and hinders users from successfully completing their tasks, e.g., an auditive medication-taking prompt while the user is on the phone. It is, thus, necessary to begin to quest for the best time to deliver interruptions caused by such services.

There are several ways for a system to present information or feedback to users, where a reminder can be considered as a usual means to prompt a piece of an instant message. We also pay our attention to the reminder service and think it as the source of the interruptions. Therefore, as the initiator of an external, purposeful interruption, reminder should be developed with the clues generated implicitly and explicitly about what users are attending to and how deeply they are focusing [Horvitz]. These clues tailored can be used to evaluate the cost and benefit of interruptions and be considered in an expected-utility decision making.

Over the last ten years, a variety of rule-based reminder systems have been explored [DeyAbowd; MarmasseSchmandt; SohnDey; Das], where a set of rules can be configured by developers [Du] or end-users [SohnDey]. A prompt, in turn, will be delivered directly only in several specific situations based on activity sensing and particular information, without

thinking at the level of a shift in user attention, particularly, when the user performs two or more activities simultaneously [Gu].

We believe that user's attention is critical in detecting if an interruption happened. The findings about how we rely on user attention in producing an unobtrusive interruption have significant implications for how we design a proper reminder service. In this paper, user attention is used to evaluate the cost and benefit of an interruption, and to make the interruption be delivered on the best time.

In this paper, we first give a brief description of interruption model and show that a desirable interruption will bring a nature transition on user's activity. We later use the activity theory [KaptelininNardi] to analyze the temporal transitions with an extension on PetriNet. The extent to which user is receptive to follow the incoming interruption and perform such a transition is then pointed out. Finally, we will summarize our results and provide a discussion about questions that still need to be explored as well as challenges for future research.

A Derived Definition and Model of Interruption

The term "interruption" is commonly used when indicating that an unanticipated task is performed [Brixey 2007]. A number of healthcare researchers have begun to study interruptions in their workplace, particularly, in nursing work [Brixey 2008][Pape 2005][Tucker 2006], making it better to understand how interruptions affect human work. We've found that the definition of interruption should be framed around an observable phenomenon relative to its source. Reminder is the most common external source that interrupts an individual with time-critical, necessary information. However, an undesirable interruption will have a negative impact on reminder performance [Speier 1999]. In order to make the individual stop his planned activity to automatically (and easily) respond to a reminder, we first clear our understanding of interruption associated with the reminder service by identifying four attributes.

Referring to Brixey's work, we focus on an interruption that is (1) a human experience for the individual who expects to be aided; (2) a means to accomplish a particular goal (e.g. health); (3) to break an organization of people's activity arising from a change of the surroundings; (4) to be handled immediately. These key attributes have been used to create the conceptual model of interruption shown in Figure 1. The model is a theoretical framework where people perform the planned activities in a sequential, interleaved or concurrent way during the pre-interruption phase. To make it clear, the term of "activity" is treated as a long-term transformation process [Li 2008] oriented toward a motive with step-by-step operations. For example, the process where people performs the concurrent activities "watch TV" and "drink" is treated as a sequence of conscious operations with timestamps, e.g., <user, living room, 8 a.m.>. Details are presented in our early work [Tang 2010].

A reminder is delivered after the Operation "OP_{x+1}" performed. The interruption phase depicts a period of time and isn't directly visible so that it is difficult to pinpoint when a person has received an interruption. Nevertheless, this stage produces effects that differ widely in user's activities. If the interruption is rejected, the planned activities continue; otherwise, the user stops to perform the prescribed activity following the instructions. A desirable interruption will bring a nature transition on user's activity.

Due to the possible negative effect of interruptions on people, it is necessary to examine interruptions in its costs and benefits. However, determining the cost is especially difficult because some of what happens is not observable [Rivera-Rodriguez 2010]. We use the

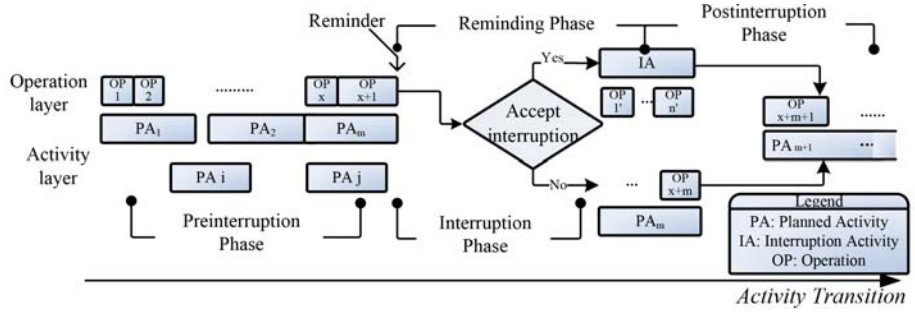


FIGURE 1: A CONCEPTUAL MODEL OF INTERRUPTION BASED ON THE ACTIVITY TRANSITIONS. THE EXECUTION OF CONCURRENT ACTIVITIES IS TREATED AS A SEQUENCE OF CONSCIOUS OPERATIONS.

observations on the change of activities to explore the clues that will make it acceptable for an interruption.

Methodology

The approach adopted focuses on seeking the best time to provide prompts with an understanding of interruptions involved into human practice. Due to the influence upon the execution of activities from interruptions, an observation towards why and how a user performs one or more specific activities enables us to point out if he is receptive to follow an incoming reminder.

Theoretical Framework for user Activity

The term "activity" is commonly used when indicating that something that people do or cause to happen. The Webster's Dictionary defines an activity as a series of "vigorous action or operation" happening at a given place and time. Over the past decade, the scholars sought to understand human activity as a complex, socially-situated phenomenon and locate activity trends with the bulk of collected sensor data. The Activity Theory (AT) was thus derived as a descriptive framework to contribute to the understanding of everyday practice in the real world, that is, "the unity of consciousness and activity" [Nardi]. With AT, the motive for an activity is created through transforming the object into an outcome [Engeström] with a sequence of conscious operations. Three defining attributes are identified.

1) Activity: A Long-term Process of Being Carried out as a Sequence of Operations

Typically, in terms of time, activity is an interval entity while operation is considered as an instant entity. In order to characterize the temporal activity, we represent a set of constraints with DAML (DARPA Agent Markup Language). An ontology of time is developed for describing two subclasses of TemporalEntity (i.e., Instant and Interval) and the relations with each other [HobbsPan]. Therefore, the time feature of interval and instant entity (In what follows, lowercase t is used for instants, uppercase T for intervals) with the first-order predicate is described as " $(\forall T)[\text{Interval}(T) \supset \text{TemporalEntity}(T)]$ " and " $(\forall t)[\text{Instant}(t) \supset \text{TemporalEntity}(t)]$ " respectively. Moreover, there are "begins" and "ends" relations on temporal entity, which can give directionality to time. Then, we can model the time-dependent feature of activity.

Definition 1 Given an interval entity T, which is described as

$$(\forall T)[\text{ProperInterval}(T) \equiv \text{Interval}(T) \wedge (\forall t_1, t_2) [ends(t_2, T) \wedge begins(t_1, T) \supset t_1 \neq t_2]],$$

then a temporal activity can be considered to be $(\forall \text{Activity}, T)[\text{during}(\text{Activity}, T) \supset \text{Interval}(T)]$, that means an activity is taking place within an interval. In the axioms, t1 and t2 are the beginning and end of interval T.

Definition 2 Given an instant entity t as $\text{Instant}(t)$, then a temporal operation is represented as

$(\forall Operation, t)[atTime(Operation, t) \supset Instant(t)]$, which means the operation is being carried out at a given time t.

The predicate "belongsTo" can be used to model the timing in relation to an activity and operation entity.

$$\text{belongsTo}(\text{Activity}, \text{Operation}) \equiv (\forall T, t)[[\text{during}(\text{Activity}, T) \wedge \text{atTime}(\text{Operation}, t)] \supset \text{inside}(t, T)]$$

Furthermore,

$$(\forall i)(\exists \text{Operation}_i)[\text{belongsTo}(\text{Activity}, \text{Operation}_i)],$$

where $(\forall t_i, t_j)[[\text{atTime}(\text{Operation}_i, t_i) \wedge \text{atTime}(\text{Operation}_j, t_j)] \supset \text{before}(t_i, t_j)]$,

then it indicates that a sequence of independent operations constitute an activity.

2) Activity: It Appears in Multiple Manners

Definition 3 A user perform activities A, B in sequential or concurrent manner. It can be represented as

$$(\forall A, B)sequential(A, B) \equiv (\forall T_1, T_2)[[\text{during}(A, T_1) \wedge \text{during}(B, T_2)] \supset \text{before}(T_1, T_2)]$$

Or

$$(\forall A, B)\,\text{concurrent}(A, B) \equiv (\forall T_1, T_2)[[\text{during}(A, T_1) \wedge \text{during}(B, T_2)] \supset \text{intDuring}(T_1, T_2)]$$

Thus,

Definition 4 Given two independent activities A and B, then the timing in relation to activity and operation entity is concluded as

$$(\forall i, A, B)(\exists \text{Operation}_i)[[\text{belongsTo}(A, \text{Operation}_i) \wedge \text{belongsTo}(B, \text{Operation}_i)] \supset \text{sequential}(A, B)]$$

Or

$$(\forall i, A, B)(\exists \text{Operation}_i)[[\text{belongsTo}(A, \text{Operation}_i) \wedge \text{belongsTo}(B, \text{Operation}_i)] \supset \text{concurrent}(A, B)]$$

then it indicates that a sequence of conscious operations groups multiple activities, which appears in a different way, especially in a concurrent manner.

We consider an interruption as a break from an organization of people's activity, for example, stop A to perform B. Thus, the findings about the transition on user activities are important for producing an unobtrusive, nature interruption.

3) Activity: it transforms due to a shift in attention to environment

AT provides us a new perspective on describing the transition to be actually caused by the reflections from the environment when a user contacts with the object world. Such environment effects (e.g., low temperature, poor light) can lead to a shift in user attention, and later initiate a specific activity. For example, with an attention on poor light, people could go to turn on the light, and then a change from the light is brought.

Based on the work by Eric Horvitz et al. [Horvitz], Figure 2 demonstrates the activities are performed due to the environmental effects on people. A set of variables (rectangular nodes) from the physical environment influence user's attention status and lead to an attention-sensitive activity which, in turn, influence the environment configuration. In other words, the environment is the precondition, as well as the post-condition of user's activity. Decisions about the computer's actions (e.g. deliver a reminder) consider balancing the costs of disruption with the value of information from a user's attention. Therefore, we consider a computer's action to be attention-sensitive.

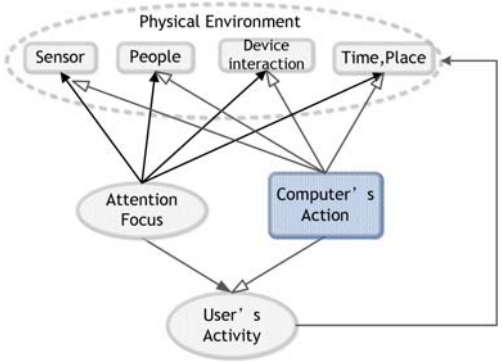


FIGURE 2: THE TRANSITION PROCESS AMONG MULTIPLE ACTIVITIES WHICH TAKES PLACE TO MEET CIRCUMSTANCES CONTINUOUSLY.

According to the interruption model in Figure 1, we believe that a desirable interruption will make a nature transition on user's activity. In order to measure the costs brought from an interruption, it is necessary to analyze the possibility or an activity or a set of concurrent activities can be broken in a particular circumstance. It is practical to characterize the possibility with Petri Net, because of its ability in describing the concurrency involved during the activity transition.

An activity is modeled as 3-tuple $\boldsymbol{N} = (\boldsymbol{S}, \boldsymbol{T}_{opt}; \boldsymbol{F})$, where,

- $\boldsymbol{S}$ is a finite set of environment states correlated to user attention, corresponding to the place in Petri Net. It emphasizes that there is a continual change in the environment within the transition process.
- $\boldsymbol{T}_{opt}$ is a finite set of discrete operations. It describes the basic elements of human activity, corresponding to the transition in Petri Net.
- $\boldsymbol{F}$ is a set of arcs,

$$F = \{I(t_1)/O(t_1), I(t_2)/O(t_2), \cdots, I(t_i)/O(t_i), \cdots\} \subseteq (S \times T_{opt}) \cup (T_{opt} \times S)\ (\mathrm{I}(\mathrm{t_i}), \mathrm{O}(\mathrm{t_i}) \subseteq S, \mathrm{t_i} \subseteq T_{opt})$$

defines the fire rules for the transition process, $\mathrm{I}(\mathrm{t_i}), \mathrm{O}(\mathrm{t_i})$ is $\mathrm{t_i}$'s input and output places respectively.

A Petri Net-based model of multi-activity is 5-tuple $\boldsymbol{N}_M$=($\boldsymbol{S}_M$, $\boldsymbol{T}_M$; $\boldsymbol{F}_M$, $\mathbf{M}$, $\boldsymbol{H}$), where,

- $S_M = \bigcup_{i=1}^{n} \{S|_{N_i}\}$, N_i denotes the i^{th} human activity,
- $\mathrm{T_M} = \bigcup_{i=1}^{n} \{\mathrm{t_i}|_{\mathrm{N_i}}\}$, t_i is the substitute for each $\boldsymbol{N}$,
- $\mathrm{M} = \bigcup_{S \in S_M} \mathrm{M(s)}$, M(S) is a number of tokens assigned to the place S ,where $\mathrm{M(s)} \geq 1$. A token can be produced if a transition t_i's preconditions (i.e. environment states) involved in its input place can be met.
- $H : S \rightarrow < \textit{human modality}, \textit{effect value} >\ (\forall S, S \in S_M)$, which models the impact on user attention from the environment in terms of different modalities, e.g., the sense of sight, auditory, or touch. A 2-tuple representation is come up.

A transition involved in the multi-activity model is enabled whenever there are sufficient tokens hold

by all input places of an activity. That is, a behavioral transition t_i can take place in a configuration of a set of environmental states P at λ moment. It is completely described via a formal syntax and semantics.

$$\exists i, t_i \in T_M, \forall p \in I(t_i)$$
$$if \quad p \in StateSet(\lambda) : M(p) \geq 1,$$
$$then \; M[t_i >$$

Figure 3 shows a composited model, including three activity models N_1, N_2, N_3 and a multi-activity model N_M, where two types of transitions occur. To N_M, the activities N_1, N_2 can be fired if and only if their preconditions are met in their input S. The execution of N_1, or N_2 consumes the tokens and produces other tokens in their output. That is, the execution would result in a continual change of environment states. The previous definitions of *sequential(A, B)*, *concurrent(A, B)* and *belongTo(activity, operation)* are shown in the multi-activity model, as it can describe different manners of performing activities. In Figure 3, the activities N_1 and N_3 are sequential, while the activities N_2 and N_3 are concurrent on the composition level. Taking multiple activities co-existed into consideration, the execution of activities is characterized as a simple time-based sequence of operations what is given in Definition 4. Then, the transfer of operations takes place due to the shifts of environment states.

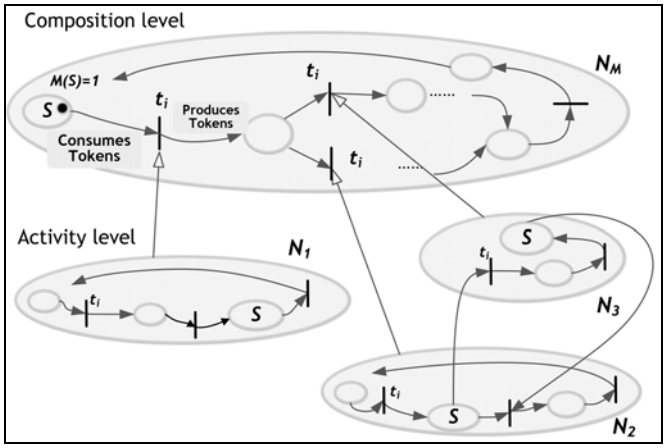


FIGURE 3: COMPOSITED ACTIVITY MODEL BASED ON PETRI NET

Figure 4 gives an example of the multi-activity model, which describes the scenario of "drinking while watching TV" involving two concurrent activities a and b. P_{ia} (i=1,2,3,4,5,6)or P_{jb} (i=1,2) is the precondition of operation t_{ia} or t_{jb}, or the post-condition of operation $t_{(i+1)a}$ or $t_{(j+1)b}$. We can find that the execution of activities is done by a sequence of operations as $\{t_1, (t_{1a}, t_{4a}) | (t_{2a}, t_{3a}, t_{4a}) | (t_{1b}, t_{2b}), t_2\}$ for example.

Decision Making on Costs of an Interruption

An interruption could happen due to a negotiation between its cost and benefit from a user. The cost is thus, related to how much attention a user will devote to breaking the activities in which he is engaged. The more a user is receptive, the less time-consuming and difficult an interruption is. There is a need to assess the cost for each incoming interruption based on activity sensing, in automated decisions about the nature and ideal timing of the appropriate reminders.

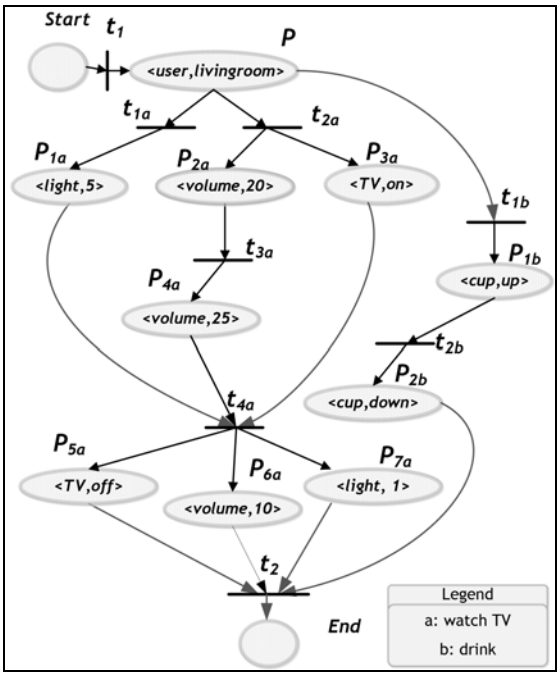


FIGURE 4: PETRI NET-BASED REPRESENTATION OF TWO CONCURRENT ACTIVITIES: WATCH TV AND DRINK

We measure the cost on user attention as the extent to which a user is receptive to follow the instructions and performs the prescribed operation, that is, the possibility of a transition of operations firing. Determining the costs is thus formulated as a reachability problem for Petri nets in our experience.

We say that a marking Mi is reachable from a initial marking M0 in one step if $M_0[t > M_i$ $(i = 1, 2, \cdots)$ when the transitions of user's operations $t \in T_M$ may continually fire in an arbitrary order in NM, where M0's set of reachable markings is the set $R(N_M) \overset{D}{=} \{M_i \mid M_0 \rightarrow^{(S_M, T_M, F_M)^*} M_i\}$.

Taking an interruption as an illustration. When the interruption is accepted, the user will stop his planned operation ti, expressed as $M_{i-1}[t_i > M_i$ $(M_{i-1} \in {}^\bullet t_i)$, and

perform the first operation tj of prescribed activity, expressed as $M_{i-1}[t_i, t_j > M_{i+1}\ (M_{i+1} \in t_j^{\bullet})$. This means that a transition is enabled in terms of the possibility of reachability, using $G(x,y)\ (x \in O(t_i),\ y \in I(t_j))$), which will in turn dictate whether an interruption could happen as a user expected. With the demonstration of the activity model in Figure 3, x, y are the post-condition and precondition of the transitions ti and ti respectively.

Inspired by Hokyum Kim et al [Kim], there are common means to get someone's attention by using sound, light, etc [Arroyo]. As a result, we structure the attention status as three perceivable modalities related to the environment, including sight, auditory and concentration, where concentration is "noticed" by specifying the strength in user's own description. Table 1 gives a general view of attention status regarding an activity being performed, expressed as <modality, H>. For example, as shown in Table 1, the constraint (i.e. preconditions) of activity "watching TV" is "user is near TV", while its influences (i.e., post-conditions) on user attentions are "sound and light intensity being up to level 7", and "user can leave less than 5 min". The measurement of perceivable modalities, denoted by H, is given in our early work [Tang].

TABLE 1 CONSTRAINTS AND INFLUENCE OF TYPICAL ACTIVITY

Activity	Constraint	Influence
Taking medicine	<scene, nearMedicine>	<dispersed, 0.52>
Watching TV	<scene, nearTV>	<auditive, level_7>; <visual, level_7>; < dispersed, 0.71>
Answering a phone	<scene, nearPhone>; <auditive, level_2>	<dispersed, 0.19>
Sleeping	<scene, onBed>; <auditive, level_0>; <visual, level_0>	<dispersed, 0>

Time for Interruption

The proposed scheme computes the possibility of an interruption happening with a decision-making on the extent to which a prescribed operation could get a user' attention away from the influences. Due to a huge number of states, an underlying problem emerged is the reduction which can be solved by data clustering as illustrated in Table 2. The states of the similar modalities form a cluster. When a state appears along with an operation, its distance measure to all known clusters is computed. Taking a new state "<TV, on, 19:30>" and the insistent state "<light, on, 19:00>" as an example, they get user's attention through human vision, in turn, the distances to the cluster are then 0.33 and 0.71. Therefore, the state "<light, on, 19:00>" is reduced.

With the c-means clustering algorithm extended, a reduction of insistent states carries out. Currently, we use a fuzzy approach for the distance computation, where the output is matched against the fuzzy sets to compute a similarity score, making it possible for a state to belong to one or more clusters. We select several concurrent activities to construct two scenarios, and there are eleven states detected in each scenario receptively.

TABLE 2 EXPERIMENT SCENARIOS AND REDUCED STATES WITH DATA CLUSTERING

Situation	States	Reduced States
Drinking while watching TV at 10:00 a.m.	<Phone, off>; <Location, living room>; <Cup, up>; <Medicine Chest, close>; <Chair, 0>; <Bed, 0>; <living room Lamp, on>; <TV, on>; <Temperature, cold>; < Noise, level_7>; <Light, level_8>	<Location, living room>; <Temperature, cold>; < Noise, level_7>; <Light, level_8>
Answering a phone during meal at 11:45 a.m.	<Phone, on>, <Location, kitchen>, <Cup, down>, <Medicine Chest, close>, <Chair, 1>, <Bed, 0>, <living room Lamp, off>, <TV, off>, <Temperature, warm>, < Noise, level_9>, <Light, level_5>	<Chair, 1>; <Temperature, warm>; < Noise, level_9>; <Light, level_5>

Now, we can consider the time for delivering a reminder, e.g., for medication taking through the use of voice under the first situation. The possibility indicating that user is receptive to follow the reminder and performs the operation, can be obtained over the reduced state space as follows. The lower-case g is the possibility over a single modality, i.e., a cluster. That means the user would like to shift his auditive attention, for example, from the current operation t_{pla} and perform the prescribed operation t_{pre} invoked by the voice reminder with a specified probability.

$$\begin{aligned}
&\forall s_m \in t_{pla}^{\bullet}, s_n \in {}^{\bullet}t_{pre}, \\
&s_m, s_n \overset{D}{=} \{ < \text{modality}, \boldsymbol{H} > | \text{modality} = \\
&\text{"visual","auditive","dispersed"} \}, \\
&g = \boldsymbol{H}\big|_{s_m} \ \Delta\, \boldsymbol{H}\big|_{s_n} \quad (\sigma > 1) \qquad (1) \\
&= \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma} exp\{ -\dfrac{(\boldsymbol{H}\big|_{s_n} - \boldsymbol{H}\big|_{s_m})^2}{2\sigma^2} \} & if\ \text{modality}\big|_{s_m} = \text{modality}\big|_{s_n} \\ 0 & if\ \text{modality}\big|_{s_m} \neq \text{modality}\big|_{s_n} \end{cases}
\end{aligned}$$

A g-related matrix D from Equation (1) is produced. Based on a single modality, M is the number of post-conditions of t_{pla}. C is the number of clusters.

$$D_C = [\, g_{CM} \,] = \begin{bmatrix} g_{11} \\ \vdots \\ g_{ij} \end{bmatrix}, \text{ where } 0 < g_{ij} < 1, i \le C, j \le M .$$

The evaluation of the cost of interruption with all attention status is based on a group of matrices. Equation (2) finds a minimum probability indicating an interruption happens overall clusters. ω_α describes the relative importance in relation to other clusters.

$$G = \min_{\alpha=1}^{C} (\omega_\alpha \times \min[g_{ij}]) \quad (2)$$

The proposed solution enables the decision-making procedure of time for an acceptable interruption. When a reminder goes off, the system recognizes a set of contextual information as states headed with simple date stamps. The clustering carries out subsequently to minimize the similarities. A match between the cost and benefit, i.e., measured by the urgency of reminder, of an interruption is negotiated to indicate if the interruption could be respective and resulting to happen currently.

Experimental results

The reminder of medication-taking is used to demonstrate that the proposed methodology can improve user adherence greatly. Research on the elderly assistance platform has centered on formulating effective principles of attention-sensitive reminder. It provides several additional health-care services, including an on-line medication-taking plan, and an easy access to the medication diary, etc. The details can be found in [Tang].

DCASI developed in our early work, accesses the flow of the contextual message from multiple sources and determines the evidence on an incoming reminder happening. The interruption detecting service performs the ongoing decision analyses with a representation on user attention. These analyses balance the urgency of reminder and the attention-sensitive costs of interruption to decide the usability of this reminder. As highlighted in Figure 5, the interruption detecting service acts as a bridge the gap between the contexts and an acceptable reminder.

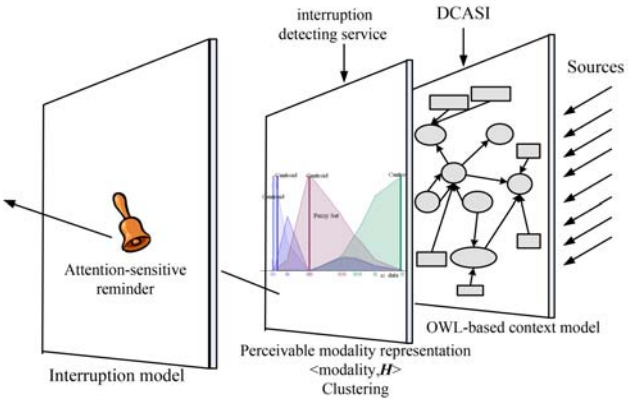


FIGURE 5: CONCEPTUAL OVERVIEW OF THE INTERRUPTION DETECTING SERVICE THAT EMPLOYS MULTIPLE MODELS TO EXECUTE AN ONGOING DECISION ANALYSIS ABOUT THE BEST TIME FOR A MEDICATION-TAKING REMINDER

We implemented the elderly assistance system, named as E-Cabinet in a real-world setting shown in Figure 6a. It serves as a user interface of the system by using a monitor pad for sensing the medication activity. The elderly only needs to pick one up, plug it in, direct it to the webcam. The E-Cabinet can then identify a new drug, plan a specific medication-taking pattern, as well as recognize the medication activity unobtrusively and automatically.

The E-Cabinet has been deployed in a real smart home where the trace collection is shown in Figure 6c. We recruited five elders over 60 years old (four males and one female) for the evaluation of our methodology in enhancing medication adherence. They were asked to use our E-Cabinet for medicine taking. For their health, they need to take a low-dose calcium tablet twice a day (one is taken with meals; the other is taken before bedtime) for three weeks. We assume the subject has taken his pills if he removes the appropriate medicine from E-Cabinet at the desired time. In order to make an ongoing comparison with time-based reminder delivery, a medication regimen, such as before bedtime, are treated as a regular pattern on time.

FIGURE 6: (A) E-CABINET WHERE TWO MEDICINES SENSED WITH RFID TAGS; (B) A PRESENTATION OF PROMPTING FOR MEDICATION-TAKING ON THE TOUCH SCREEN, IT SAYS "DOSAGE: 3 PILLS A DAY BEFORE MEALS";(C) ROOM WITH LIGHT, NOISE SENSOR, RFID LOCATED SYSTEM.

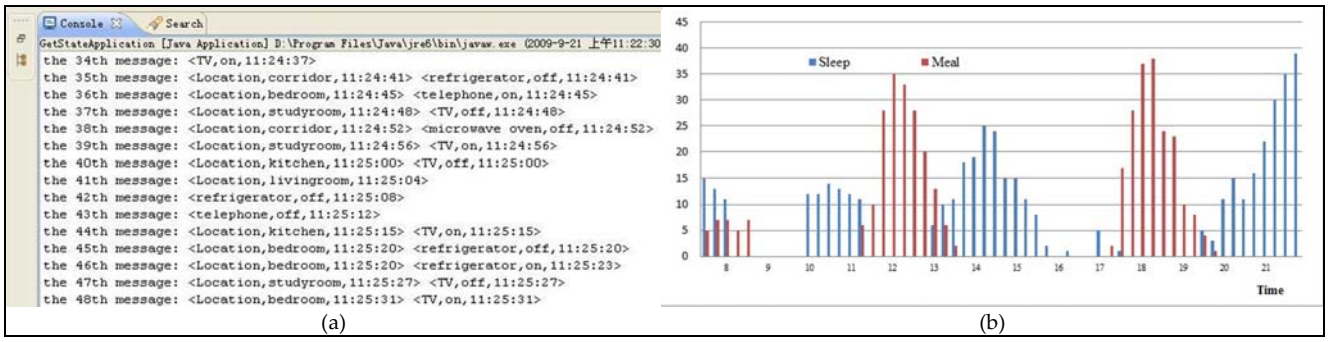


FIGURE 7: (A) TRACKING RESULT OF ADLS (ACTIVITIES OF DAILY LIVING); (B) FREQUENCY OF OCCURRENCE OF "SLEEPING" AND "HAVING A MEAL".

By tracking the user's activity, we got the digested regular patterns as follows, "Having a meal should begin between 11:30 a.m. and 12:15 a.m." and "Sleeping after 21:00 p.m.". Accordingly, the time-based reminders are produced as "to prompt for taking tablet at 12:00 a.m. and 21:00 p.m.". Figure 7 shows the tracking result of subjects' daily activities with RFID and sensors on the appliances.

Survey responses confirm the E-Cabinet's value beyond the attention-sensitive prompting. We now move on to the research question. Similar to Pamela J. Ludford et al's [Ludford] work of user study, as Figure 8 illustrates, we depict data from the surveys compared with above time-based reminder delivery. For a given reminder delivery, we asked subjects:

(a) Whether the reminder delivery disturbed them.

(b) Whether they respond to the delivery while they were performing certain activities

(c) Whether the reminder helped them remember to do the medication-taking task.

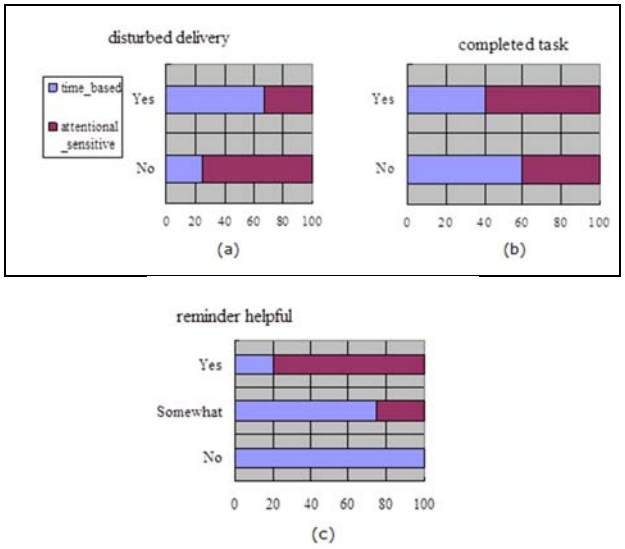


FIGURE 8: SURVEY RESPONSES SHOW SUBJECTS ARE MORE LIKELY TO ACCEPT, COMPLETE TASK AND FIND REMINDERS HELPFUL IF THEIR ATTENTIONS ARE GOT APPROPRIATELY.

We adopted the mean adherence rates to make a comprehensive assessment of system performance with two measures. The "dose taking adherence" represents whether all tablets are taken in a day, and the "dose time adherence" indicates whether the tablets are taken according to the particular dosage pattern (i.e., with meal and before bedtime). Table 3 shows the results. The attention-sensitive prompting resulted in a significantly better adherence (90.1%) as compared to the time-based prompting (87.3%) and non-prompting (75.8%). In addition, we observed that the adherence (e.g. 44.7% using non-prompting) where the subject was asked to take medication according to the specific dosage pattern was consistently worse than the adherence (e.g. 75.8%) without any requirements. The adherence (e.g. 61.0% using time-based prompting) was better, but the subject reported that he/she would miss the prompting when he/she was on the lookout for other things. The difference under the attention-sensitive prompting was much less and the adherence (79.6%) is enhanced.

TABLE 3 MEASURES OF ADHERENCE

Measures	Non-prompting (%)	Time-based prompting (%)	Attention-sensitive prompting (%)
Dose taking adherence	75.8	87.3	90.1
Dose time adherence	44.7	61.0	79.6

We constructed the situation as shown in Table 2 to evaluate the interruption detecting service. When a subject was answering a phone, a medication-taking reminder was hold until the subject finished his call, and a voice-based one was then delivered on the subject's phone. The comparisons on a decision-making about the best time with between the rule-based and our approach can be found in our early work [Tang]. The former reasoning process was quite computationally intensive. In order to make a reminder acceptable, it is important to anticipate all

possible contextual situations and assert the contexts into the context knowledge base. It, however, does at a great expense on the execution time because it has to examine whether each case in a rule is true. In addition, the rule-based way would fail when the subjects are engaging in two or more activities. In our experiment, when the situation "answering a phone during dinner" occurs, the reminder will fail because it cannot describe the "answering" and "eating" simultaneously, that would brings a conflict on the reasoning. Our method can generate an effective reminder because it evaluates the reminder based upon an understanding of the acceptable interruption with user attention. As a result, it is possible to improve user's adherence to a reminder effectively.

Discussion

We often generate clues implicitly and explicitly about the time for delivering a reminder. However, a person who is focusing deeply has trouble with adherence, because when to initiate the followed fluid interactions relies mainly on whether his attentions are got appropriately. We try to bridge the gap between human attention and contextual clues around the world. In the light of activity theory, the reflections on changing environment occur along with an acceptable reminder. To our knowledge, it is the first empirical study that has a shift of user attention considered to happen under a sequence of prescribed operations. The more likely user performs an operation, the better chance a reminder has of being receptive.

Activity theory is a theory of object-driven activity [Engeström]. Traditionally, it considers an entire activity system beyond just one actor. We make clever use of its structure and bring the environment as an external actor to the system. A user performs the operations to produce certain outcomes which in turn change the environment states and resulting a shift in attentions. We only explain how social environment as well as artifacts involved mediates the user's activity within a system and observes the formation of activity transition. A limitation is lack of object-oriented structure which we intend to support in the future, making it possible to build a shared object between two or more systems.

The goal of this study is to examine user acceptance to an incoming interruption brought by a reminder. Although the interruption examined in this study has been investigated for some notification systems, this study does provide insights on user attentions. The finding showed that someone's attentions can be got and quantified by using the perceived influences related to the surroundings on his senses. It suggests that a reminder may be more respective and effective if the required environmental constraints of a specific operation were considered to be met under the current influences. This study is based on the attention status separated as only three common types of human modalities. However, there are limitations with regard to capturing a shift in user attentions on smell, heat, etc. Future studies may be able to evaluate user acceptance from a more systematic view designed to recognize the individual differences and the changing environmental conditions.

As this study was only conducted with a small set of users, the scalability of the system needs to be measured. A larger study with twenty-five elder users over three months is in the planning. Compared with a typical rule-based reasoning method, our proposed approach will have a great improvement in terms of running time, since it focuses on the physical effects on human modalities for the measurement of the efficacy of a remind.

Related Work

Situation-aware reminder Delivery. Providing a reminder service at the right time is gaining a great deal of attention from the research community. It is easy to think of an appropriate reminder being discovered under an understanding of user situation. We often generate clues implicitly and explicitly about what we are selectively performing in a certain environment. Place-Its [Sohn] runs on the mobile phones to study people using reminders. It enables the integration of the location-based reminders into people's daily practice. As such, location is commonly used as the means for getting someone's situational information. Early proof-of-concept designs include CybreMinder. It has developed a fully featured reminder application based on Context Toolkit, as well as been built using a wired hardware assembly available. The work, including Dede [Jung] and Gate Reminder [Kim], defined the basic idea – a reminder is configured at design time, and delivered at the appropriate location. Similar to Place-Its, PlaceMail [Ludford] supports the useful reminders and functional place-based lists to work indirectly towards when and where to deliver the location-based information. People's daily plans of activities have been used to coordinate the reminder delivery recently. The Autominder System [Pollack] is designed to assist

individuals with cognitive impairment in ADL completion. It produces the reminders by a decision-making on if there are the discrepancies between the planned and observed activity. These AI-based planning reminders are dedicated to guide when errors are made during task completion for multiple-step ADLs. Das and colleagues [Das] used the machine learning techniques to determine when to deliver prompts for different IADLs (instrumental ADLs) in a smart environment. Although these approaches are effective for predetermined tasks, it doesn't do a very good job of resolving the conflicts between the reminder delivery and unplanned situation fluidly and quickly, as well as improving the individual's positive experience.

Adherence to a reminder. The behavioral change is most likely to occur for adequate adherence when the cost-benefit analysis of using prompting supports is positive, and the user is satisfied with the outcome [Seelye]. HYCARE [Du] coordinates the reminder delivery by handling the possible conflicts from the asynchronous events happening while a prompt is being delivered. It allows for some unplanned situations from strict time and event-based schedule and balancing the costs of disruption with the value of urgent prompts. It was a good try to add the flexibility of the prompting system to everyday life. From the perspective of user, there are obviously still conflicts between his planned and prescribed actions as shown in Figure 1. The cost of promoting comes from how much attention he will devote to breaking the activities in which he is engaged. As of time of writing, there has not been done much work on it, let alone in-depth research, however. We guide the research towards examining when to pass an acceptable reminder and expressing the solutions by inspecting the reminder delivery via a spotlight of attention. Attention cues are central in the decisions about when to initiate or to make an effective contribution to a conversation [ClarkSchaefer] or project, e.g. Attentional User Interface (AUI) project [Horvitz]. More generally, detecting or inferring the attention is an essential component to the process of grounding—converging in reasoning about the situation awareness versus the disruption of users [Horvitz]. In our methodology, by structuring the user's attention status as the three perceivable modalities of environment, we have pursued the use of attention cues as an indicator of user acceptance, and forecasted the future cost brought from an incoming reminder.

Interruption. The growing interest of reminders in persuading the users to accept the instructions is a welcome development. Such attractive services have turned out to be the interruptions upon the users. Observational studies in the workplace show that the recipients of an interruption have not the response in the 40% of the cases [O'ConaillFrohlich]. Thus, seeking the right time for interruptions is a challenging task to improve user adherence. It says that once the system knows the user's activities and his specific operational structures [Chávez], as well as the dependencies between the operations and environment, it may be able to predict the most probable interruption on the user in this situation. Although there is a rich history of prior work on the interruptions from the human-computer interfaces [GillieBroadbent] [Cutrell] [McFarlane], we have found that there is much we do not yet understand, especially when to interrupt the users in a multitasking environment where people often perform multiple activities at the same time. In this work, we study the interruption framework and explore the clues that make it acceptable for an interruption with the observations on the activity transitions. The algorithm has been developed to compute the probability over breaking an or a set of concurrent activities towards an incoming reminder.

Conclusion

We suggest the ways to minimize the disturbance of interruptions for resolving the effective delivery issue. To get the users' attentions to conduct an acceptable interruption, we use the observations on the change of activities to explore the attention clues. Experimental setting around a medication-taking situation is built as an illustration for the efficiency of proposed solution in improving user adherence. When it comes to reminder delivery, we believe our study is just the beginning. Further research can uncover the additional parameters that impact an effective delivery. However, our study finds that human attention can be used to negotiate a match between the cost and benefit of an interruption. Furthermore, the activity theory is suited to support broadly for inferring the best delivery point. Our results take us closer to a future where the attention-sensitive reminders will improve the way people complete their tasks.

ACKNOWLEDGMENT

This work is partially funded by National Natural Science Foundation of China (No. 61303041), The Project Sponsored by the Scientific Research Foundation for the Returned Overseas Chinese

Scholars in Shaanxi Province of China.

REFERENCES

Arroyo, E., T. Selker, et al. (2002). Interruptions as multimodal outputs: which are the less disruptive? Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on, IEEE.

Chávez, E., R. Ide, et al. (1999). "Interactive applications of personal situation-aware assistants." Computers & Graphics 23(6): 903-915.

Clark, H. H. and E. F. Schaefer (1987). "Collaborating on contributions to conversations." Language and cognitive processes 2(1): 19-41.

Cutrell, E., M. Czerwinski, et al. (2001). "Notification, disruption, and memory: Effects of messaging interruptions on memory and performance."

Das, B., C. Chen, et al. (2010). Automated prompting in a smart home environment. Data Mining Workshops (ICDMW), 2010 IEEE International Conference on, IEEE.

Dey, A. and G. Abowd (2000). CybreMinder: A context-aware system for supporting reminders. Handheld and Ubiquitous Computing, Springer.

Drews, F. A. (2007). The frequency and impact of task interruptions in the ICU. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, SAGE Publications.

Du, K., D. Zhang, et al. (2008). Handling activity conflicts in reminding system for elders with dementia. FGCN (2) 2008, IEEE: 416-421.

Du, K., D. Zhang, et al. (2008). "HYCARE: A hybrid context-aware reminding framework for elders with mild dementia." Smart Homes and Health Telematics: 9-17.

Engeström, Y. (2009). "The future of activity theory: A rough draft." Learning and expanding with activity theory: 303-328.

Engeström, Y., R. Miettinen, et al. (1999). Perspectives on activity theory, Cambridge University Press.

Gillie, T. and D. Broadbent (1989). "What makes interruptions disruptive? A study of length, similarity, and complexity." Psychological Research 50(4): 243-250.

Gu, T., Z. Wu, et al. (2009). epsicar: An emerging patterns based approach to sequential, interleaved and concurrent activity recognition. PerCOM 2009, IEEE: 1-9.

Hobbs, J. R. and F. Pan (2004). "An ontology of time for the semantic web." ACM Transactions on Asian Language Information Processing (TALIP) 3(1): 66-85.

Horvitz, E., A. Jacobs, et al. (1999). Attention-sensitive alerting. Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc.

Horvitz, E., C. Kadie, et al. (2003). "Models of attention in computing and communication: from principles to applications." Communications of the ACM 46(3): 52-59.

Jung, Y., P. Persson, et al. (2005). DeDe: design and evaluation of a context-enhanced mobile messaging system. Proceedings of the SIGCHI conference on Human factors in computing systems, ACM.

Kaptelinin, V. and B. A. Nardi (2006). Acting with technology, Mit Press.

Kim, H., I. Park, et al. (2007). Dynamic Activity Lifecycle Management in Ubiquitous Computing Environments. CIT 2007: 985-990.

Kim, S. W., M. C. Kim, et al. (2004). Gate reminder: a design case of a smart reminder. Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques, ACM.

Ludford, P. J., D. Frankowski, et al. (2006). Because I carry my cell phone anyway: functional location-based reminder applications. Proceedings of the SIGCHI conference on Human Factors in computing systems, ACM.

Marmasse, N. and C. Schmandt (2000). Location-aware information delivery with commotion. Handheld and Ubiquitous Computing, Springer.

McFarlane, D. C. (2002). "Comparison of four primary methods for coordinating the interruption of people in human-computer interaction." Human-Computer Interaction 17(1): 63-139.

Nardi, B. A. (1995). Context and consciousness: activity theory and human-computer interaction, mit Press.

O'Conaill, B. and D. Frohlich (1995). Timespace in the workplace: Dealing with interruptions. Conference companion on Human factors in computing systems, ACM.

Pollack, M. E., L. Brown, et al. (2003). "Autominder: An intelligent cognitive orthotic system for people with

memory impairment." Robotics and Autonomous Systems 44(3-4): 273-282.

Rivera-Rodriguez, A. and B. T. Karsh (2010). "Interruptions and distractions in healthcare: review and reappraisal." Quality and Safety in Health Care 19(4): 304-312.

Seelye, A., M. Schmitter-Edgecombe, et al. (2011). "Application of Cognitive Rehabilitation Theory to the Development of Smart Prompting Technologies."

Sohn, T. and A. Dey (2003). iCAP: an informal tool for interactive prototyping of context-aware applications. CHI Extended Abstracts 2003, ACM: 974-975.

Sohn, T. and A. K. Dey (2004). iCAP: Rapid Prototyping of Context-Aware Applications. Proceedings of CHI, Citeseer.

Sohn, T., K. Li, et al. (2005). "Place-its: A study of location-based reminders on mobile phones." UbiComp 2005: Ubiquitous Computing: 903-903.

Tang, L., X. Zhou, et al. (2011). "MHS:a Multimedia System for Improving Medication Adherence in Elderly Care(accepted)." IEEE System Journal.

Tang, L., X. Zhou, et al. (2010). Adaptive prompting based on Petri Net in a Smart medication system. PerCom Workshops 2010, IEEE: 328-333.

Weiser, M. (1991). "The computer for the 21st century." Scientific American 265(3): 94-104.

Lei TANG (Sichuan, China, 1983) received the PhD degree in computer science and technology in 2012.

She is currently work in the school of information engineering, Chang'an University, P. R.China.

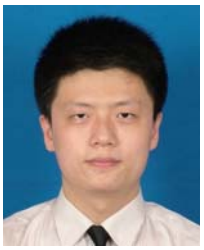
She has been a visiting researcher at the chair of information systems, Mannheim University, Germany in 2009-2010. Her research interests include the area of pervasive computing and user-oriented computing.

Dr. Tang is a member of ACM, IEEE, and a member of CCF (China Computer Federation)

Zongtao Duan (ShaanXi,China, 1977) is currently a associate professor at the school of information engineering, Chang'an University, P. R. China. He received his Ph.D. in computer science from Northwestern Polytechnical University, P. R. China in 2006.

He has been a postdoctoral research fellow in the University of North Carolina, American in 2009-2010. His research interests include context-aware computing in transportation.

Dr. Duan is a member of CCF, CCF High Performance Computing, and Pervasive Computing Technical Committee.



Hanbo Wang (ShaanXi, China, 1983) received the PhD degree in computer science and technology in 2012.

He has been a visiting researcher in echnische universitt münchen, Germany in 2009-2010. His research interests include the area of pervasive computing in electric power and embedded systems.

Zhiwen YU (Hubei, China, 1979) is currently a professor at the school of computer science, Northwestern Polytechnical University, P. R. China. He received his Ph.D degree of Engineering in computer science and technology in 2005 from the Northwestern Polytechnical University, China.

His research interests cover pervasive computing, context-aware systems, human-computer interaction.

Dr. Yu is a member of ACM, IEEE, and IEEE Computer Society, a senior member of CCF and CCF Pervasive Computing Technical Committee.

International Journal of Computer Science and Application (IJCSA) Volume 3 Issue 2, May 2014 www.ij-csa.org
doi: 10.14355/ijcsa.2014.0302.04

Location Based Security and Message Authentication Services in Vehicular Ad Hoc Network (VANET)

P.Chandrasekar[1], Arul Lawrence Selvakumar[2]

[1]Assistant Professor, Dept. of Computer Applications, S. A. Engineering College, Chennai, India.

[2] Professor & Head, Dept. of CSE, Rajiv Gandhi Institute of Technology, Bangalore-32, India.

[1]mail2chandruu@yahoo.co.in; [2]aarul72@hotmail.com

Received 28 December 2013; Accepted 15 January 2014; Published 15 May 2014

Abstract

The most recent years have observed an unmistakable convergence of Vehicular Ad hoc Networks (VANETs) that promise to modernize the way we drive. Information about the location is fundamental objective in Vehicular Ad-hoc Networks. Nearly, all VANETs rely on location based information, addressing security and privacy issues as the requirement of VANETs development; it's difficult to avoid any possible malicious attacks and resources abuse. Employing a digital signature scheme is widely recognized as the most effective approach for VANETs to achieve authentication, integrity, and validity. The proposed system exposes the location based security and message authentication services in vehicular ad hoc networks.

Keywords

VANETs; Security and Privacy Issues; Message Authentication

Introduction

According to Traffic Safety Facts Annual Report from the National Highway Traffic Safety Administration, nearly 6 million police-reported motor vehicle crashes occurred in the United States alone in 2006, leading to 1.75 million injuries and 38,588 deaths. According to the 2006 Annual Report on Traffic Congestion in the Denver Region, each resident on average faces about 32 hours of congestion delay per year. Travel during rush hours is 27% longer than non-rush hours. $1.7 billion per year is lost due to the traffic delays. The above numbers indicate that the traditional traffic crash alert and traffic control systems should be meliorated in order to improve the quality of the public transportation. Fortunately, various manufacturers, government units and standardized bodies have generated national and international association committed exclusively to VANETs.

Vehicular Ad-hoc Networks (VANETs) have appeared as a new appliance that is envisioned to modernize the human driving experiences, optimize traffic flow control systems, etc. Location is fundamental information in Vehicular Ad-hoc Networks (VANETs) and Addressing security and privacy issues as the prerequisite of VANETs' development must be emphasized. Almost all VANET applications rely on location information. Therefore it is very importantee to make sure location information integrity, meaning that location information is original, secure, correct (Not false or fabricated) and unmodified (value not changed). According to the security and privacy it is difficult to avoid any possible malicious attacks and resources sharing abuses and employing various digital strategy scheme is widely recognized, whereas, this is the most effective approach for VANETs to achieve message authentication, integrity and secure validity.

The number of information received by a vehicle becomes large; a scalability problem emerges immediately, where a vehicle could be difficult to sequentially verify each time received data within 100-300 ms interval in accordance with the current Dedicated Short Range Communications (DSRC) Protocol. Nowadays we have a vehicle tracking system using GPS is available but due to some natural obstacles we couldn't track it on the hills and valley sides. Vehicles and roadside infrastructures are equipped with wireless communication devices and

constitute a vehicular ad hoc network (VANET). VANET aims at improving the road safety and avoid potential traffic accidents. This technology mostly helps us to improve not only the country growth, but also saves the human value from the timing incidents.

Smart and Sophisticated Vehicles' in VANETs

In vehicular network the new technology brings not only updated design but also new devices to vehicles. These new devices can extend vehicle's capabilities in computing, communication and sensing. The provision of on-board Global Positioning System (GPS) devices has modernized driving. Similarly, the recent introduction of short-range radar in some top-of-the-line models promises to reduce the number of fender-benders, quick-struck and other accidents. In fact, on-board radar has already been used in advanced cruise control systems. We base our vehicle model on the smart vehicle proposed by Hubaux et al. Hubaux includes an Event Data Recorder (EDR), a GPS receiver, and radar in his smart vehicle model. It is equipped with various features such as wireless transceiver, Digital Map, Common Service Centre, GPS receiver, radar communication, a location key identification and etc.

Location Information

VANET is composed of vehicles and road side units (RSUs). Vehicles are equipped with wireless communication devices, which are called on-board units (OBUs). The wireless communication devices enable vehicles to exchange traffic related messages with each other and with RSUs. The location in a message can be part of the message as shown in Figure 1a. E-Business applications for VANETs (online shopping, toll payment collection, etc.) can use this type message to enhance the security of messages. The whole content of a message can be composed of location information in OBUs, shown in Figure 1b. Military vehicle movement on the arena can use this type of message. The location information of all vehicles is aggregated and it can be strictly encrypted and decrypted in a specified decryption region to enhance the security of the location information. We also expect the decryption region can move along with the receiver vehicle.

Data at OBU	
Address	Location

(a)

Address
Location-1
Location-2
……
Location-n

(b)

FIG. 1: TWO TYPES OF LOCATION MESSAGES: (a) LOCATION IS PART OF THE MESSAGE AND (b) LOCATION IS THE WHOLE CONTENT OF THE MESSAGE

Location information can be obtained from several devices, including GPS receiver, infrared scanner, etc. Since GPS imposes some constraints such as lack of other positioning techniques have been proposed for the vehicular field, including cellular or WiFi localization, dead reckoning (by using last known last location information and velocity), ultrasound range sensors and image/video localization.

Message security and privacy issues in VANETs

Addressing the security and privacy issues in VANETs includes message verification and conditional privacy preservation. Nowadays, Vehicles have been equipped with more and more high-technology devices. For example: GPS System, Radars and OBUs. This type of wireless enabled devices makes vehicle intelligent and be able to "talk" with each other, and thereby form a self-organized VANET. With the assistance of Vehicle to Vehicle communications, potentially fatal road accidents can be avoided; dangerous driving behaviors can be alerted; city traffic flows can be optimized; traffic jams can be alleviated. However, even though VANETs bring tremendous benefits to us, VANETs raise many research challenges as well. One of these challenges is security concerns. In VANETs, malicious vehicles may modify or insert fake information in the network, which could incur life-endangering accidents. In a word, if the security mechanism in VANETs is not carefully designed, misbehavior and malicious attacks may ruin the original intention of VANETs. Therefore, prior to putting VANETs into the practice, it is important to have a robust and efficient security mechanism on board.

Security Threats in Location Based Privacy VA-Network

Since security and privacy are basically important in VANETs, recently more and more research efforts have been put on designing security and privacy preservation protocols for location based privacy. In

VANETs, there are several possible security threats are available in location based privacy VA-network. Such that *False information attack, DoS attack, Replay attack, Impersonation attack, Message modification attack, Privacy attack and Trajectory disclosure attack in a particular location.* Generally, to escape from these attacks in the location based VANET that the following five requirements are directly associated with the security threads. Such are authentication, confidentiality, integrity, conditional privacy, and scalability. The implication of message authentication and security related schemes are motivated by this fact that, the proposed system introduces an efficient RSU-Aided Message Authentication scheme named RAIMA, for VANETs.

RAIMA explores an important feature of VANETs by employing RSUs to assist vehicles in authenticating messages. With RAIMA, vehicles first perform mutual authentication and key agreement with an RSU. Vehicles that received safety messages do not need to verify the message through a conventional PKI-based scheme. Instead, each safety message will be attached with a short Message Authentication Code (MAC) that is generated by a sender under the secret key shared between the sender and an RSU. The major contribution of RAIMA improves the authentication efficiency and reduces the communication overhead.

Design Objectives

The importance to ensure the security of location information has not been seriously addressed until recently. PKI is an important way to ensure information confidentiality. We assume there is a trusted authority. The trusted authority will generate keys for PKI. PKI has two keys, one is a public key which will be known by everyone and the other is a private key which is known only by the owner. The PKI algorithm is known to all parties. If a public key is used by the PKI algorithm to encrypt information, a cipher text will be generated. PKI is naturally adopted in VANET.

RSU-aided Message Authentication Scheme Using PKI Algm.

Towards a better understanding of RAIMA, it is important to understand the following four steps as in Figure 2. Which are *Registration, Key Establishment, Hash Aggregation, and Verification.* When vehicles enter the communication range of an RSU, vehicles initiate a mutual authentication process with an RSU. An RSU authenticates vehicles by verifying their signatures. Vehicles compute message signatures with their private keys, and RSUs verify the signatures and their corresponding public key certificates. A valid signature means that the signer of the signature is a legitimate user in VANETs. In a similar way, vehicles can also authenticate an RSU. Meanwhile, the messages that have been signed by vehicles and RSUs include secret credentials which can be used to compute a PKI used shared key. Here, Diffie-Hellman key agreement could be adopted to establish the shared symmetric *key*. It is worth noticing that different vehicles share different keys with an RSU. Vehicles do not know the key shared between an RSU and other vehicles.

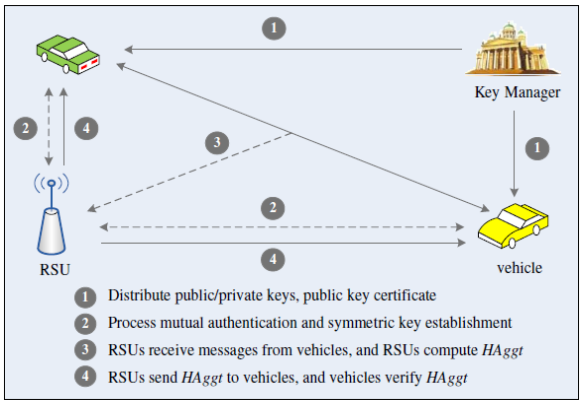


FIG. 2: THE ILLUSTRATION OF THE LOCATION BASED MESSAGE COMMUNICATION USING RAIMA

PKI brings new challenges to VANETs. PKI includes both the public key and the private key. The public key needs to be known by other vehicles and the private key needs to be secretly stored. Therefore, key management is one of the challenges because of the large scale of the vehicle population. Some vehicles may have expired keys. To update the public and private keys, pervasive infrastructure will be required. Another challenge is that RSU message authentication scheme using PKI that requires homogenous configuration to achieve communication among vehicles. Some old vehicles may not have PKI installed. In addition, the overhead of PKI in terms of processing time will add significant processing time overhead to VANETs as well.

RAIMA Key Establishment

When a vehicle *Vi* detects the existence of an RSU *R* (e.g., through a Hello message from the *R*), the *Vi* initiates anonymous mutual authentication and establishes a shared secret key with the *R*. This can be achieved by adopting the Diffie-Hellman key establishment protocol secured with signature scheme. The mutual authentication and key establishment

processes are shown as follows:

$$Vi \rightarrow R : \{aP \mid CertVi\}PKR.$$
$$R \rightarrow Vi : IDi \mid bP \mid \{IDi \mid aP \mid bP\}SKR.$$
$$Vi \rightarrow R : \{IDR \mid bP \mid aP\}SKVi$$

If the above three steps are completed correctly, the mutual authentication succeeds. Note that the mutual authentication in the protocol is provably secure (refer to for more details).

Validating Locations and Message Communication

We simulated a bidirectional 3 km highway with two lanes in each direction. The cell radius was 100 meters, traffic arrival rate was 1600 vehicles/hour, mean velocity was 33.3 m/s, and transmission radius initially was 100 meters. Each simulation has some number of compromised, or malicious, vehicles and a single observer vehicle. When the observer enters the simulated highway, it initiates a request to find all of the compromised vehicles. The simulation terminates when the observer reaches the end of the 3 km highway. To determine how transmission range affects the time needed to detect malicious vehicles, we ran a set of experiments with a 100 m transmission range and a set with a 500 m transmission range. The vehicle density was about 30 vehicles per kilometer per lane on the highway. The compromised vehicles were 5% of the total vehicles and were randomly deployed along the highway. In each set of simulations, we varied the length of the highway to investigate the effect of transmission range.

Finally, TABLE 1: General Integrity Case: Parameters and Values shows and to investigate how many compromised vehicles could remain undetected in our system. We randomly distributed compromised vehicles along the highway. The vehicle density was about 30 vehicles per km on the highway.

TABLE 1: GENERAL INTEGRITY CASE: PARAMETERS AND VALUES

Parameters	Values
Initial traffic density	30 vehicles/Km/lane
The length of the road L	3 Km
Average speed	60 km/h
The number of lanes	4/direction
The mean error μ	1 m
The deviation of error σ	1 m
Error ϵ	3 m
# of neighbor outliers m_n	4
# of opposite outliers m_o	1
The weight for radar w_1	0.5
The weight for opposite w_2	0.3
The weight for neighbors w_3	0.2

The transmission radius was 100 meters. We varied the percentage of all vehicles that were compromised from 0% to 30% because our assumption is that the majority of vehicles are honest. Compared with previous compromised vehicle percentage (5%), we increased this number to 30% to investigate a larger scope. We measured the number of compromised vehicles detected in 30 seconds.

Conclusion and future work

Vehicles installed with advanced devices can communicate with each other and create wireless networks, called VANETs. The applications of VANET include safety and entertainment applications. Most, if not all, of these applications are location-aware and strongly depend on location information. Compared with other networks, such as MANETs, the Internet, and cellular networks, VANETs have unique and distinctive features. We are protecting a tuple, correct time, correct identity, and correct location. Based on information security requirements, we have to ensure location information confidentiality, integrity and availability. The research in this dissertation addresses location information security by providing location information confidentiality, integrity, and availability.

To ensure inter-cell location integrity, the aggregated position information of several vehicles is transmitted over the wireless medium which is open to the public. If the aggregated message is in plaintext, it is vulnerable to an assortment of attacks. One simple solution is to encrypt the plaintext message by using conventional cryptography. However, key management is a very challenging task given the huge number of participating vehicles. This is my future work to follow along with a geographic location-based security mechanism for the providence of physical security on top of conventional methods. The security and privacy message transfer needs to encrypt messages with a secret key which is converted from geographic location information.

REFERENCES

A.Wasef, and X.Shen, "ASIC: Aggregate Signatures and Certificates Verification Scheme for Vehicular Networks", in *Proceedings of IEEE Global Communications Conference (GLOBECOM '09)*, Nov, 2009.

A.Wasef, and X.Shen, "MAAC: Message Authentication Acceleration Protocol for Vehicular Ad Hoc Networks",

in *Proceedings of IEEE Global Communications Conference (GLOBECOM '09)*, Nov, 2009.

A.Wasef, and X.Shen,"REP: Location Privacy for VANETs Using Random Encryption Periods", *ACM Mobile Networks and Applications (MONET)*, to appear.

C.Adams and S.Lloyd, Understanding PKI: Concepts, Standards, and De-ployment Considerations. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2002.

Dedicated Short Range Communications (DSRC), [Online]. Available: http://grouper.ieee.org/groups/scc32/dsrc/index.html.

Douglas R.Stingson, "Cryptography: Theory and Practice, third edition," *CRC Press*, 2005.

G.Calandriello, P.Papadimitratos, A.Lioy, and J.P.Hubaux, "Efficient and robust pseudonymous authentication in VANET", in *Proceedings of the International workshop on Vehicular ad hoc networks (VANET'07)*, 2007.

J.H. Song, V. W. S. Wong and V. C. M. Leung, "Wireless Location Privacy Protection in Vehicular Ad-Hoc Networks", *Mobile Networks and Applications*, 2009.

J.P. Hubaux, S. Capkun, and J. Luo, "The security and privacy of smart Vehicles", IEEE Security and Privacy Magazine, vol. 2, no. 3, pp. 49-55, 2004.

P.Bahl and V.N.Padmanabhan, "Radar: A in-building RF-based user location and tracking system", in Proceedings of the IEEE Infocom, Tel Aviv, Israel, MAR 2000, pp. 775-784.

R.Lu, X.Lin, and X.Shen, "SPRING: A Social-based Privacy-preserving Packet Forwarding Protocol for Vehicular Delay Tolerant Networks", in p*rocessings of the IEEE International Conference on Computer Communications (INFOCOM'10)*, Mar, 2010.

R.Mobus, M.Baotic, and M.Morari, "Multi-Object Adaptive Cruise Control", Hybrid Systems: Computation and Control, vol. 2623, pp. 359-374, Apr. 2003.

S.Dietzel, E.Schoch, F.Kargl, B.Konings, and M. Weber, "Resilient Secure Aggregation for Vehicular Networks", *IEEE Network - Special Issue on "Advances in Vehicular Communications Networks*, 2010".

S.Spors, R.Rabenstein, and N.Strobel, "A multi-sensor object localization System", in VMV '01: Proceedings of the Vision Modeling and Visualization Conference 2001. Aka GmbH, 2001, pp. 19-26.

W.Diffie and M.E.Hellman, "New Directions in Cryptography", *IEEE Transactions on Information Theory*, Vol. 22, No. 5, pp. 644-654, 1976.

Y.Sun, X. Lin, R. Lu, X. Shen, and J. Su, "A Secure and Efficient Revocation Scheme for Anonymous Vehicular Communications", in *Proceedings of IEEE International Conference on Communications (ICC'10)*, May, 2010.

Y.Sun, X.Lin, R.Lu, X.Shen, and J.Su, "Roadside Units Deployment for Efficient Short-time Certificate Updating in VANETs", in *Proceedings of IEEE International Conference on Communications (ICC'10)*, May, 2010.

Y.Xi, K. Sha, W. Shi, L. Schwiebert, and T. Zhang, "Enforcing Privacy Using Symmetric Random Key-set in Vehicular Networks", in *Proceedings of the International Symposium on Autonomous Decentralized Systems (ISADS)*, pp. 344-351, 2007.