A Perception Model of Spam Risk Assessment Inspired by Danger Theory of Artificial Immune Systems

Kamahazira Zainala, Mohd Zalisham Jali

Universiti Sains Islam Malaysia (USIM), Nilai, Negeri Sembilan, Malaysia

Abstract

This present paper relates Danger Theory of Artificial Immune Systems, which has been introduced by Polly Matzinger in 1994 with the application in risk assessment. As to relate the concept of Danger Theory in risk assessment, a situation of determination severity level for detected Short Messaging Service (SMS) spam is applied. However, further testing is needed as to demonstrate the explained concept. Danger Model that based on the idea of the immune system is appear to be suitable as the fundamental principles and the most generic available solution as to assimilate its theory into the risk assessment environment especially that involve severe or hazardous impacts.

1. Introduction

Much research has been done that inspired from the human immunological theories. The research covers many areas that have contributed a high, effective impact such as in the field of chemical spectrum recognition, web data mining, robotics, network and computer security and dynamic learning (Dasgupta, 2006). This paper describes the application of one of the immunology theory, which is Artificial Immune System (AIS). Specifically, the scope of concern is about the link of Danger Theory applied in risk assessment of information security and also as the underpinning theory of this paper.

Liu et al., (2010) applied AIS in their proposed model to evaluate risk for information system security. However, they are not mentioning any specific immunity principles or AIS algorithm used. In this paper, the mathematical model where the detectors recognized threats and its’ intensity is developed. In addition to AIS, they used Analytic
Hierarchy Process (AHP) to measure the severity level which is the function of damage degree for confidentiality, integrity and availability (CIA).

In Zhuang et al., (2009), a model of Danger Theory based information security Risk Assessment (DT-RA) is proposed. This model that specifies Danger Theory to analyse risk claimed that the service capacity and efficiency of the assessment are greatly improved.

This paper is arranged as following: Section 2 explains the idea of risk assessment in information security, with Section 3 describes the Danger Theory of AIS. Then, Section 4 elaborates the relation of Danger Theory in risk assessment with the proposed framework to rank the severity of detected SMS spam, as an illustration of Danger Theory application. Further discussion and conclusion are elaborated in the subsequent section.

2. Risk Assessment

According to the National Institute of Standards and Technology (NIST, 2012), risk is related to the assessment of the extent to which an entity is threatened by a potential circumstance or event where the potential circumstance is directly related to the likelihood of threats occurrence. While, information security risks are those risks that arise from the loss of confidentiality, integrity, or availability (CIA) of information or information systems and reflect the potential adverse impacts to organizational operations, organizational assets, individuals, other organizations, and the nation. Risk in information security basically associated with CIA value of the information itself.

\[ R = f (C, I, A) \]  

Basically, risk assessment is the critical phase which the evaluation of risk will influence the decision on how to deal with it, this means to consider the right treatment, in order to response to identified risk by planning its risk treatment and to minimize its possible hazardous impact.

There are a few steps involved in the management of risk. These steps fundamentally consist of four phases: identification, assessment, response or treatment and monitoring of risks. This paper is further cover only the part of risk assessment or risk evaluation which this phase measures the degree impact of detected risk. The degree of risk will be translated into three levels of prioritized severity which are high, medium and low. The process of evaluation is possibly involved either qualitatively or quantitatively.

3. Danger Theory of AIS

AIS, which is a type of Computational Intelligence (CI) model is originally inspired from the human immunology. This Danger Theory was found by Matzinger (2002) which suggests that the immune system able to recognise invaders and reacts to such threats based on the correlation of various (danger) signals.

There has been a wide usage and contribution of Danger Theory especially in anomaly detection such as intrusion (Lu, T et al., 2012) and virus detection (Zekri, M et al., 2014). Dasgupta (2006) summarized the development of AIS in computer security and fault detection.

In human immunology, two types of protection are:
- Innate immune system (divided into first line and second line of nonspecific defence mechanism); and
- Adaptive immune system (specific defence mechanism which consist of humoral and cell-mediated immunity).

Dendritic cell which is an innate immune system is truly an intrusion detection agent of the human body. They are Antigen Presenting Cell (APC) that process antigen material and present it on the cell surface to the T cells of the immune system. They act as messengers between the innate and the adaptive immune systems.

This specific dendritic cell function has inspired the idea of Dendritic Cell Algorithm (DCA) (Matzinger, 2002). The idea of Danger Theory is formed with these functions:
- danger signals will be stimulated by foreign invaders that are dangerous via initiating cell stress or death.
- immunology system manipulates these danger signals to recognise the danger zone and then evaluates the danger.
**Input**: $S = \text{set of data items to be labelled safe or dangerous}$  
**Output**: $L = \text{set of data items labelled safe or dangerous}$

```plaintext
Begin
    Create an initial population of dendritic cells (DCs), $D$
    Create a set to contain migrated DCs, $M$
    for all data items in $S$ do
        Create a set DCs randomly sampled from $D$, $P$
        forall DCs in $P$ do
            Add data item to DCs’ collected list
            Update danger, PAMP and safe signal concentrations
            Update concentrations of output cytokines
            Migrate dendritic cell from $D$ to $M$ and create a new DC in $D$ if concentration of costimulatory molecules is above a threshold
        end
    end
    forall data items in $S$ do
        Calculate number of times data item is presented by a mature DC and semi-mature DC
        Label data item as safe if presented by more semi-mature DCs than mature DCs otherwise label as dangerous
        Add data item to labelled set $M$
    end
end
```

Fig. 1. Generic Danger Theory algorithm.

Pereira (2011)\(^8\) stated that when a cell is injured, it emits a distress signal and activates the local APCs which consequently co-stimulate Helper T cells, and these in turn help B cells that finally activate the immune response.

Immature dendritic cell in tissue constantly samples the surrounding environment for any possible pathogen such as viruses and bacteria. Once they have come into contact with a presentable antigen, they become activated into mature dendritic cell and begin to migrate to the lymph node, where B and T cells are stored for further immunological action. Immature dendritic cell phagocyte pathogens and degrade their proteins into small pieces and upon maturation present those fragments at their cell surface using Major Histocompatibility Complex (MHC) molecules.

Green smith et al., (2010)\(^9\) stated that the signals migrated to the lymph node are divided into two types of signals regarding to the degree of detected danger. Apoptotic alerts or semi-mature brings the safe signal, while necrotic alerts brings the mature signal. Semi-mature implies a ‘safe’ context and mature implies a ‘dangerous’ context. These signals are reflection of the state of the surrounding. The mature signals are distributed into two forms, which is Pathogen Associated Molecular Patterns (PAMP) signal that indicate abnormal behaviour which is highly indicate of an anomaly. The other one is the danger signal that is less severe than PAMPs. A general sign of system distress called as inflammatory signals. All of these signals can be depicted as in Fig. 5(a).

The collected data of antigen by dendritic cell is measured using the Mature Context Antigen Value (MCAV) which this is the mean value of context per antigen type. This can be measured by the following equation:

$$MCAV(\text{antigen\_type}) = \frac{\text{mature\_count}}{\text{antigen\_count}}$$  

(2)

The closer the value of MCAV to 1.0, the higher probability of ‘context’ (antigen) type being correlated with the danger signals.

The Danger Theory takes care of “non-self but harmless” and “self but harmful” invaders into our system. The central idea is that the immune system does not respond to non-self but to danger. Practically there is no need to attack everything that is foreign. In this theory, danger is measured by damaging or injuries to cells that is indicated by distress signals that are sent out when cells die in an unnatural death.
According to Greensmith et al., (2010), Danger Theory model also can be viewed as an extension of Two-Signal model, where these two signals are antigen recognition (signal 1) and co-stimulation (signal 2). Co-stimulation signal is used to indicate the corresponding antigen is dangerous.

In addition to that, Greensmith et al., (2010) also elaborated that lymphocyte behaviours in this theory are based on three laws:

- **Law 1:** A lymphocyte will be activated if the system receives both signal 1 and signal 2. If the system only receives signal 1 without signal 2, then an activated lymphocyte will die. If the system only receives signal 2 without signal 1, then it ignores the signal.
- **Law 2:** A lymphocyte only accepts signal 2 from APC. Any cells can issue signal 1 to the system. An experienced T-cell or B-cell can act as APC.
- **Law 3:** After the activation of a lymphocyte, it will be reverted to the resting state after a while.

### 4. Mapping of Danger Theory into Spam Risk Assessment

Spam is one type of threat that could lead to hazardous situation when the likelihood of potential risk to occur is high. Spam is considered as a way to exploit vulnerabilities. In this case, vulnerability is a flaw or weakness in a system which referred to human weakness as the main factor and vulnerabilities in mobile phone and computers as the secondary factors. The impact or risky situation will take place when the probability of vulnerabilities to be exploiting via this threat is high.

SMS spam management (Sulaiman et al., 2014) basically consist of three phases which are spam detection, classification and severity determination level. In this concept paper, the scope of security risk assessment is only cover the final phase in SMS spam management, that is the determination of risk or severity level. This is taken as risk assessment example to describe the measurement of danger signal density in the content of SMS.

Risk is a measure of the extent to which an entity is threatened by a potential circumstance or event, and is typically a function of:

- the adverse impacts that would arise if the circumstance or event occurs; and
- the likelihood of occurrence.

With reference to Blank et al. (2012) that produced a guidance document about conducting risk assessments under National Institute of Standards and Technology (NIST), risk assessments are a key part of effective risk management to facilitate decision making. Risk management processes include:

- identification of risk;
- assessing risk;
- responding to risk; and
- risk monitoring.

Risk models vary in the level of point of risk concentration and complexity with which threat events are identified. As in spam management, there are three main processes involve which is

- spam classification;
- spam clustering, which detected spam is categorized into groups according to its subject-matter; and
- determination level of spam’s severity which also include the options available to respond against spam.

The spam management model is depicted in Fig. 2.
The concept between Danger Theory and determination of SMS spam severity level is mapped in Table 1.

Table 1. Concept mapping between immune system model, Danger Theory and SMS spam.

<table>
<thead>
<tr>
<th>Danger Theory property and type of signals</th>
<th>Biological property</th>
<th>Abstract property in risk assessment of detected SMS spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antigen</td>
<td>Bacteria, pathological cells</td>
<td>SMS spam or threat</td>
</tr>
<tr>
<td>Dendritic cell</td>
<td>Collecting signals around damaged cells</td>
<td>SMS spam folder</td>
</tr>
<tr>
<td>PAMP signals</td>
<td>PAMPs alerts – indicate the presence of microbial</td>
<td>High risk – signature of likely anomaly</td>
</tr>
<tr>
<td>Danger signals</td>
<td>Necrotic alerts – indicate the damage of tissue</td>
<td>Medium risk – indicate potential anomaly</td>
</tr>
<tr>
<td>Safe signals</td>
<td>Apoptotic alerts – indicate of healthy tissue</td>
<td>Low risk – indicate an absence of anomaly</td>
</tr>
<tr>
<td>Context</td>
<td>Antigen Presenting Cells</td>
<td>Content of SMS spam</td>
</tr>
<tr>
<td>Inflammation</td>
<td>Inflammatory cytokines – indicate general tissue distress</td>
<td>Confirmation of SMS is a spam</td>
</tr>
<tr>
<td>MCAV</td>
<td>Information collection by dendritic cells</td>
<td>Term frequency (that is considered as spam) in specific category</td>
</tr>
<tr>
<td>Antigen storage</td>
<td>-</td>
<td>Spam database library</td>
</tr>
<tr>
<td>Semi-mature signal</td>
<td>represents the dendritic cell of likely safe</td>
<td>Low risk – indicate an absence of anomaly</td>
</tr>
<tr>
<td>Mature signal</td>
<td>represents the dendritic cell of likely danger</td>
<td>High risk – signature of likely anomaly; or</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Medium risk – indicate potential anomaly</td>
</tr>
</tbody>
</table>
This proposed model is depicted in Fig. 4. The scope of Danger Theory application is highlighted in coloured box.

![Diagram for proposed Danger Theory inspired in assessing SMS spam severity level.](image-url)

The details of Fig. 4 are explained in Table 2, which also it’s integration relationship with Fig.2.

Table 2. Mapping of relationship between Fig. 2 and Fig. 4.

<table>
<thead>
<tr>
<th>Spam management model</th>
<th>Proposed model for spam severity determination level</th>
<th>Justification</th>
</tr>
</thead>
</table>
| Classification        | Phase 1                                             | As in spam management model, classification of spam messages is the initial phase by identifying whether it is legitimate or spam. In previous research, this detection of spam messages can be executed using various techniques such as machine learning algorithms, feature and content selection and many more. The detected spam will be clustered according to the respective group or category. The sample of spam category such as competition, free prizes, advertisement, financial assistance will determine its risk level according to the impact loss of identified threat. The phase of determination of spam’s severity level will be divided into three main stages:  
  - High – spam event could be expected to have a severe or catastrophic adverse effect  
  - Medium – spam event could be expected to have a serious adverse effect  
  - Low – spam event could be expected to have a limited or negligible adverse effect  
| Determination of Severity Level | Phase 3                                             | The action taken to response is according to recipient’s discretion either to ignore, delete or even escalate the spam message to authority body. The tagged of spam message with high and medium risk labelled indirectly has suggest to recipient the possible action to response. |
| Response               | Not Applicable in this figure                       |               |
In Section 3, MCAV is defined as the mean value of context per antigen type, or in SMS spam environment, this value is the frequency of possible spam term existence accordingly to category that finally will rank the severity. This ranking is reflecting the degree of impact or loss.

As in a common risk assessment, the risk level of a threat is measured via function of impact and likelihood of identified threat exploiting vulnerabilities, which this is widely used.

<table>
<thead>
<tr>
<th>Impact</th>
<th>Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Medium</td>
<td>Low</td>
</tr>
<tr>
<td>High</td>
<td>Medium</td>
</tr>
</tbody>
</table>

The risk is prioritized according to its possible impact such as high risk will cause a chaotic in a country if recipients believe in with the spam content. Risk is considered as medium level if the content would lead recipients to some money loss issue. And the low level of risk would be considered as equal as to ‘safe’ situation or will not bring any hazardous condition.

![Diagram](image)

(a) The transformation between signals in dendritic cells (DC)

(b) The assessment of risk level in spam messages

Fig. 5. The transformation between signals in dendritic cells (DC) and its application in spam risk assessment.

This rank can be achieved via MCAV that basically calculated as the degree of the impact once the SMS spam getting high probability of trustworthy from the recipients for further action towards the received SMS. As in addition to complement the utilization of MCAV, AHP will be further applied as to quantify the impact value (high, medium, risk).
In Liu et al., (2010)² the researchers proposed the use of AHP to evaluate the severity, the damage degree of the confidentiality, integrity and availability.

AHP is a combination of qualitative and quantitative analysis method on decision making which is proposed by Thomas L. Saaty. AHP has been used as one of the known method in risk evaluation.

Tsai et al., (2011)¹² used AHP to assess the existence of risk in wireless networks which purposely to assist administrator in monitoring the security of wireless network. The hierarchy in this paper is developed from the perspectives of the risk severity, security requirements, attacks and the configurations involved in wireless network setup.

In other research, Ji et al., (2010)¹³ applied AHP in security assessment and weight verification of live application environment which comprises of a server, firewall and four units of workstations as clients. As a result, AHP methodology in network security assessment is proved able to improve the accuracy and efficiency of network risk assessment.

4.1. Initial Version of Algorithm

With reference to generic DCA in Fig. 1, here is the initial version of the algorithm. The designed algorithm is based on Fig. 4 of proposed Danger Theory inspired in detection of SMS spam severity determination level.

In this algorithm, input is referring to the incoming messages that already classified as spam, while output is the expected outcome of this algorithm where the input (spam messages) are finally tagged with the associate risk level that the spam message carried with.

The calculation of risk level for every spam messages will be further defined as to enhance this algorithm. This calculation is essentially reflects the MCAV measurement that is introduced in the Danger Theory. As stated in above paragraph, AHP will be used as to quantify the degree of severity.

This initial version of algorithm will be further enhanced to measure the hazardous level of identified spam. Risk assessment technique will be integrated with this algorithm in order to design and develop an enhanced version of existing Danger Theory. An enhanced version of this theory will be applied in this field.

```plaintext
Input :  S = set of spam messages to be labelled as high, medium or low risk
Output: L = set of spam messages labelled as high, medium or low risk

Begin

identified Risk \rightarrow
Create a database of spam with risk indicator, D (database library)
Create a folder to contain risk-labelled spam, M (folder for spam with risk level indicator)
for all spam messages, S in P do
    Create a set of spam messages from sample in D, P (spam folder)
    for all spam messages in P do
        Add data item in D
        Update information on high, medium and low risk level
        Update risk level of spam cluster with related spam term
        Migrate S from P to M and create a new data item in D, if current information not available
        S then become L
    End
End

for all L in M do
    Label L as to be high, medium or low risk
End

assessed risk level
for all spam messages, S do
    Calculate the risk value accordingly with the spam content
    Label spam messages with the high, medium or low risk
    Add spam messages with risk level indicator into M
End

Prioritize risk
End
```

Fig. 6. Initial version of spam risk-labelled algorithm based on DCA which is the focal point is to determine the severity level of detected spam.
The calculation of risk intensity of spam messages is possible to be implemented using the methodology of AHP. This is depicted in Fig. 7 and justified in the following paragraph.

As mentioned in Section 2, risk or impact loss is basically associated with CIA value of the information itself as formulated in equation (1). The level of risk is characterized by CIA value and will be further assessed in the lower layer by utilizing MCAV in equation (2). In the end, every spam message is labeled according to its assessed risk; high, medium or low.
5. Conclusion

Presently, there are quite a number of simple applications for spam filtering that are available on the Internet and can be installed in the smartphone. These applications are capable to distinguish between valid and spam messages which will be kept in a different folders. But these applications’ capability only limited to differentiate messages into ham and spam, and no further information on the detected spam. As a result, users still tend to open the spam messages and respond inappropriately, especially to the captivating one.

As this research will study the determination of risk level for every detected spam, the similar idea could be use in email system environment. Since, till at this moment of writing there is no published research has been found in the area of severity determination or spam prioritization level for detected email spam. This risk assessment on spam, which appropriate level of severity will be tagged to spam and this could assist users in making decision on how to respond against spam accordingly by understanding the possible concentration of unfavorable impact.

With reference to previous research papers in other fields of risk assessment, the applied Danger Theory concepts in risk evaluation is expected to give results of optimum level in accuracy and efficiency in risk evaluation and severity ranking. Other than that, indirectly this research is expected to contribute findings in text or data mining.

Acknowledgements

Authors wish to thank the Ministry of Education (MOE), Malaysia for funding this research. The research grant is known as RAGS, under the code number RAGS/2013/USIM/ICT04/2.

References