

## A Review of Predictive Analytic Applications of Bayesian Network

Mohammad Hafiz Mohd Yusof<sup>#</sup>, Mohd Rosmadi Mokhtar<sup>\*</sup>

<sup>#</sup>Research Center for Software Technology and Management, Faculty of Information Science and Technology,  
Universiti Kebangsaan Malaysia, Bangi, Selangor, 43600, Malaysia  
E-mail: hafiz.yusof@siswa.ukm.edu.my, mrm@ukm.edu.my

**Abstract**— Malware can be defined as malicious software that infiltrates a network and computer host in a variety of ways, from software flaws to social engineering. Due to the polymorphic and stealth nature of malware attacks, a signature-based analysis that is done statically is no longer sufficient to solve such a problem. Therefore, a behavioral or anomalous analysis will provide a more dynamic approach for the solution. However, recent studies have shown that current behavioral methods at the network-level have several issues such as the inability to predict zero-day attacks, high-level assumptions, non-inferential analysis and performance issues. Other than performance issues, this study has identified common scientific characteristics which are reduced parameter,  $\theta$  and lack of *priori* information  $p(\theta)$  that causes the problems. Previous methods were proposed to address the problem, however, were still unable to resolve the stated scientific hitches. Due to the shortcomings, the Bayesian Network in terms of its probabilistic modeling would be the best method to deal with the stated scientific glitches which also have been proven in the area of Clinical Expert Systems, Artificial Intelligence, and Pattern Recognition. This study will critically review the predictive analytic applications of Bayesian Network model in different research domain such as Clinical Expert Systems, Artificial Intelligence, and Pattern Recognition and discover any potential approach available in the domain of Computer Networks. Based on the review, this paper has identified several Bayesian Network properties which have been used to overcome the abovementioned problems. Those properties will be applied in future studies to model the Behavioral Malware Predictive Analytics.

**Keywords**— malware analysis; behavioural analysis; Bayesian Network

### I. INTRODUCTION

Recent studies have shown that current behavioural methods at the network-level have several issues such as the inability to predict zero-day attacks, high-level assumptions, non-inferential analysis and performance issues [1].

The reduction of millions of features, disregarded parameters, removed similarities of most of the traffic flows to reduce information noise, limited number of features and ignores instances which are not entity are amongst others have been identified as the main issues contributing to the inability to predict zero-day attacks. Meanwhile, the assumed suspicious connection to be larger, longer and seldom larger, high-level features of the size of infected, recovered or removed hosts, assumption on neighbour state or user activity, assumption on malware messages' features have contributed to the issue of high-level assumption, notwithstanding the issue of data frequency, ratio, min and median threshold value, and percentages are attributed to describe the descriptive analysis on the collection of main features which have contributed to the issue of non-inferential analysis and finally excessively detailed features, too many modelled parameters, rule-based, resource-

consuming of logic and algorithm have contributed to the performance issues.

This research is intended to further expound on the three issues of inability to predict zero-day attacks, high-level assumptions, and non-inferential analysis by focusing on the specific technical issues and the shortcomings of the existing proposed solution. The research will leave the discussion on performance issues on the different specific research paper as it is essential to consider that performance is closely related to protocols' issues in different environments such as UDP and TCP [2] and performance differentials are accessed through varying network load and mobility [3].

### II. MATERIALS AND METHODS

Based on the above-mentioned issues, those problems could be further grouped into their mutually shared scientific characteristics or common criteria which are summarized in the following points:

#### A. Common Criteria

1) *Reduced Parameters,  $\theta$* : Numerical characteristic of a population is often denoted by parameter  $\theta$  and numerical description of a subset is denoted by  $y$  which both is

uncertain before a dataset is obtained and the level of uncertainty decreases once the dataset is identified. Given space  $\Theta$  is the set of possible parameter values  $\theta$ , thus  $\theta \in \Theta$  [4] so that the product of all possible outcomes of parameter  $\Theta$  and unknown parameters  $X$  becomes  $\Omega$  denotes the universal,  $\Omega=X \cdot \Theta$  [5] thus it is important to obtain as many information about the parameters as possible to derive informative results.

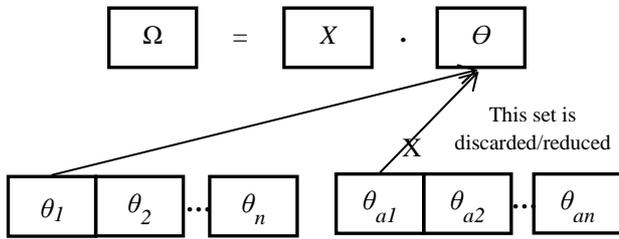


Fig. 1 Reduced parameters explained in diagram

As illustrated in Fig. 1 above, conceptually, these are the building block of a universal set  $\Omega$  which is the outcomes of all possible parameters and the unknown parameters. Given  $\Theta$  is the space of all possible parameter values  $\theta$  where  $\theta \in \Theta$ . In the diagram, there are two sets of parameters  $\theta_n$  and  $\theta_{an}$ . These sets are the element of the parameter space  $\Theta$ . If let say a method is used to reduce or discard each of the parameters set, it will limit the parameter information which could be used to draw further conclusions or the inability to predict unknown (zero days) attacks. This problem could be solved through prior information.

Rahbarinia, B et al. [6] used pruning rules, for instance, a query rule of  $\leq 5$  domains, queries  $\geq 99.99$  percentile, population, query  $\geq 1/3 \theta_m$  meanwhile hereby Zaman, M et al. [7] applied whenever there is  $i$ , such that  $t_i \leq t$  &  $t_{i+1} > t$  and the  $i$  is the single parameter  $\theta$  of the extraction feature of *http* which both approaches are limiting the adequate feature information. Edem and Feizollah [8, 9] used the K-mean as the method to reduced noise in feature information as for instance suppose there are two instances of  $M$  and  $N$  attributed to the coordinated of two parameters of  $X_{ij}$  and  $Y_{ij}$  in an initial centroid value of  $C_1$  and  $C_2$ . The distance is calculated using Euclidean distance,  $D$  as below

$$D = \sqrt{\sum_{i=1}^n \sum_{j=1}^n \|x_i - x_j\|^2 + \|y_i - y_j\|^2} \quad (1)$$

The process is repeated and  $X_{ij}$  and  $Y_{ij}$  will be grouped again into similar groups based on the new minimum distance to the centroid. This clustering technique tends to ignore the instances that are not the entity of any of the formed clusters thus reduce the parameter,  $\theta$  information.

Studies from Zaman, Edem and Feizollah [7]-[9] also reveal that previous methods and current behavioural research are highly dependent on instances or sign to feed into the feature selection process. Meanwhile, Villalba and O'kane [10], [11] used  $n$ -gram of the given  $n$  size of instructions  $p=I_1, I_2 \dots I_n$  and, from a  $N=2$  program structure that composed of two opcodes of one or more operands, the operands are then discarded left only the set of opcodes of  $o: p=o_1, o_2 \dots o_n$ . This discarded step will reduce the parameters.

2) *Lack of Priori*,  $p(\theta)$ : Prior distribution  $p(\theta)$  describes our belief that  $\theta$  represents the true population characteristics [5]. As shown in Fig. 2, For instance, research in Wen, S et al. [12] applied state transition which usually depends only on the current state to determine the outcome of the next state and does not take into consideration the outcome of the previous state [13].

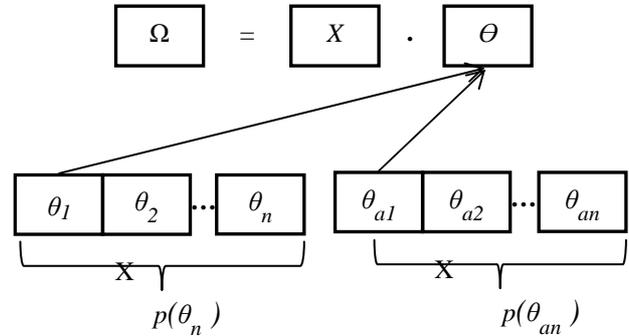


Fig. 2 Lack of priori explained in diagram

This is not the case with intrusion and malware detection, whereby any of the previous states or transactions is taken into account. In state transition, Markovian chains, for example, every new stage or the outcome at any stage is called current state. For example, give the following transition:

$$P_{ij} = P(X_{t+1} = j | X_t = i) = P(X_{t+1} = j | X_0 = i) \quad (2)$$

The transition shows above only depends only on going one step ahead. It is always  $X_t = i = 0$  and  $X_{t+1} = j = 1$ , where the state will go from 0 to 1. This will not describe the characteristic of prior knowledge of any event. The limited *priori* leads to limited conclusion to be drawn.

Ahmad, Xue and Arora [14]-[16] used ratio, percentage, frequency, average distribution to represent the collection of information of the main features in data collection without determining inferential or in-depth analysis thus lack of prior knowledge,  $p(\theta)$  [17]. In the future, this ratio could be applied with additional traffic ratio between application layer protocols with further refinement and representation through Poisson distribution model to enable the generalisation of traffic behaviours throughout the research domain [14].

Xue, et al. [15] assumed the suspicious connection denoted by  $SP$  to be larger and of longer duration, which led to the conclusion that whenever a connection packet is more than or equal to ten,  $n \geq 10$ , and the duration exceeds 10 minutes, that connection is considered malicious. This could be true in some instances; however, detailed experimental data should be provided to support such an argument. Prior information is formed using probabilities.

Another example of is the study of Wen, S et al. [12], who assumed that the state of neighbouring nodes is independent and that neighbouring nodes are in the same broadcast domain, although in the real production network

they are not necessarily in the same collision domain. This assumption shows it lack of *priori* information.

In a volatile or in a critical infrastructure network environment such as in energy industry, the lack or prior information could cause catastrophic false alarm as happened in the history of Iranian nuclear plant and in Saudi Aramco oil and gas plant.

3) *Liaison between Issues and Common Criteria*: Fig. 3 shows the logical relationship between highlighted issues and the common criteria of the scientific hitches mentioned in subsection A and B. Reduced parameter has been identified causes the inability to predict zero-day attacks mainly because answers of the uncertainty of any distributions or data observations are obtained from how much finite amount of information contained within the data at hand [18]. Parameters  $\theta$  is a numerical characteristic of a given population or space  $\Theta$  of which often signifies as  $\theta \in \Theta$  [4]. This is the building block of the whole universal set  $\Omega$ , thus obtaining the maximum amount of parameter information is a paramount task for which failure will lead to false alarm and incapability of interpreting future attacks.

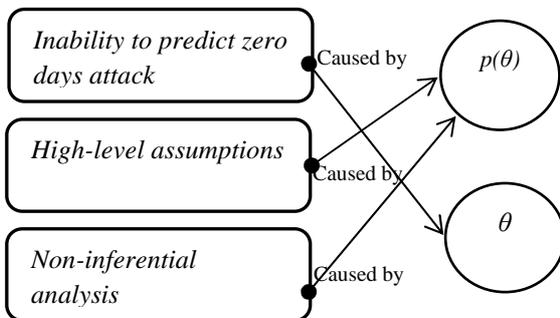


Fig. 3 Logical relationship between issues and common criteria

Meanwhile, lack of *priori* information causes the high-level assumptions and non-inferential analysis. Prior distribution  $p(\theta)$  describes the preceding scientific belief system that  $\theta$  represents the true population characteristics [5] and this gives some sort of probability function which could establish the randomness or uncertainty associated with the parameter  $\theta$  which in turn determine not necessarily the total outcome but also the consequences of the experiment [18]. This could be done by determining the probability function of the parameter at certain values. By contrast, high-level assumptions and non-inferential analysis will never go through such discipline because of the descriptive nature of its analysis where it just represents the collection of information of the main features in data collection without determining inferential or in-depth analysis [17].

Before going any further, some introduction of the background studies of this research is explained here. This paper is the continuation of a few published papers on behavioural analysis which is focusing more on network level analysis. Paper Yusof et al [37] discusses an overview of functionality of malware analysis studies of system and network level from various literature in terms of technical aspects of signature and behavioural based analysis methods which includes Support Vector Machine, Poisson, Rule-

based, Set-Function, Longest Common Sequence, K-Mean Clustering, Relationship Functions and open source which generally categorised as machine learning, statistical and tool based analysis method. Feature selection and evaluation technique in host and network level are also summarized in this paper. Overall in this paper, the epistemological aspect of this research domain is underlined and introduced to the research community with the formation of general malware analysis research framework and a few useful tables that tabulate information with regards to methods previously used in system and network level malware analysis, tabulation of feature selection and evaluation techniques. This paper established the trending area in this research domain which is the behavioural based method which could drive a future researcher to look into this area of studies.

Paper Yusof, M. H. M et al [1] meanwhile further details out the nomenclature of behavioural analysis methods specifically in the area of computer networks. As a continuation of paper [37], it is further confirmed that signature-based analysis that is done statically is no longer sufficient to solve malicious attacks problem, therefore, a behavioural or anomalous analysis will provide a more dynamic approach for the solution. This paper has critically and intensively reviewed more literature especially in the area of behavioural malware analysis studies in the computer networks. It reveals a few shortcomings of previous analysis methods and established a discussion on why Bayesian Network is preferably the best method to cope with the stated problems.

In parallel to that, this paper is designed to introduce the readers to some of the important interdisciplinary topics surround Bayesian Network method more specifically in the area of Clinical Expert Systems, Artificial Intelligence, and Pattern Recognition.

Since the malware analysis in computer networks, in general, are less studied due to the lack of leveraging behaviour of the malware attack in the network environment as mentioned by Nari et al. [38], this paper realised that behavioural analysis research in network level is still a new domain which applies approaches that might have simple knowledge based or statistical approach used to address the scientific hitches, due to that this paper discussed some sophisticated statistical approach that suits in complexes and volatile environment and proven across different disciplines.

### B. Related Method

A few methods have been introduced to resolve problems of the shared characteristics of the common criteria from the previous discussion. The solution is basically coming from the current trends in malware analysis method which has been identified as in the class of probability theorem, fuzzy, statistical analysis and clustering [1]. The solutions are rule-based, correlation statistics and state-transition of Markov Chain.

1) *Method Overcomes Reduced Parameters  $\theta$* : One of the well-known methods for knowledge manipulation and knowledge representation was a rule-based system which has been applied by Zaman and Petel [7], [19] in the form of  $R_1$ : if  $\theta_1$  then  $\theta_2$  where the  $\theta_2$  statement of consequence can be determined with level of certainty whenever  $\theta_1$  statement of condition is observed. Let another rule conditions that  $R_2$ :

if  $\theta_2$  then  $\theta_3$  where  $\theta_3$  statement represents the forward chaining factor which involves the rules of  $R_1$  and  $R_2$  immediately after  $x_1$  is established in the chain.

Note that such rules are unbalanced or asymmetric in the sense that the statement of condition and the statement of consequence are not interchangeable (switchable), such that by observing the statement of consequence does not allow us to conclude the statement of the condition [20]. Say, some event  $m$  is known to cause the effect in event  $n$  and the relationship of both events are known to be deterministic. Hence, the causal relationship between  $m$  and  $n$  can be formulated as a rule such that "if  $m$  then  $n$ " rather than "if  $n$  then  $m$ ".

Rule-based system ignores some mechanisms or causal parameter of "causal direction" which makes this method is assured only up to certain level of precision. For instance, in medical expert systems scenario consider the causal chain of smoking causes bronchitis causes dyspnoea which is denoted by the following concatenation rules as introduced in [20]  $R_3$ : if smoking then bronchitis and  $R_4$ : if bronchitis then dyspnoea. Bronchitis is a respiratory infection which is the main airways of the lungs (bronchi) is inflamed and becomes irritated whilst dyspnoea is a medical term for shortness of breath a symptom of bronchitis [21]. Now, let assume that the  $R_4$  is formulated as  $R_4$  rule which states  $R'_4$ : if dyspnoea then bronchitis. This statement would make smoking and dyspnoea become contending conditions for bronchitis which consequently would not be able to determine the condition of the patient's breathing patterns. Obviously, the rule based is inappropriate for representing the nature of causal relations amongst events [20] which in turn Bayesian consider every possible rule in the form of *priori* and *posteriori* information to represent the causal relations.

K-means by Edem and Feizollah [8], [9] then clusters  $n$  (the parameters) into  $K$  clusters around centroid  $C$ , where  $C = \{C_1, C_2, \dots, C_k\}$  given  $S = \{S_1, S_2, \dots, S_k\}$ , number of partitions. The process of iteration to get new centroid is used to overcome the reduction of parameters; however, this iteration process is based on the ratio between two points. This will further distance the plausible parameter information as at the first place is has been reduced then the iteration makes the plausible information distance even further which finally will draw wrong information.

2) *Method Overcomes Lack of Priori  $p(\theta)$* : Correlation statistics methods as in relational function and average distribution have been closely introduced by Arora, Ahmad, Wen and Xue [12], [14]-[16]. Correlation is a statistical association between two events which often infers or implies causal even if there is not a direct connection relation between the two parameters events. A correlation may signify the occurrence of hidden parameters which are common causes of the observed events, thus makes them statistically associated. Take for instance in Fig. 4 that illustrates the example below of whether smoking denoted by  $S$  causes bronchitis denoted by  $B$  or whether there exist additional veiled parameters denoted by  $\theta$  that cause both events which have been introduced by [5].

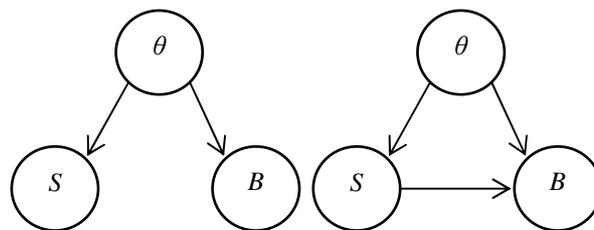


Fig. 4 Correlation between smoking,  $S$  and bronchitis,  $B$  and the hidden parameters  $\theta$

In a randomized experiment, it is a practice that whenever there is a possibility of unknown hidden parameters, it is necessary to separate the "cause" in order to conclude that there exists a causal relationship which is achieved by a controlled or randomized experiment. For randomized experiment, each level of treatment groups is chosen randomly.

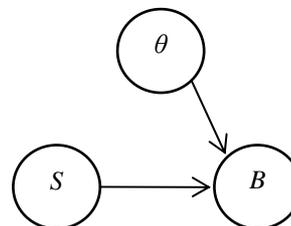


Fig. 5 Correlation between smoking,  $S$  parameters  $\theta$  and bronchitis,  $B$  in a controlled experiment

Meanwhile, like in the case of Fig. 5, group of smokers and non-smokers are assigned randomly to carry out "controlled experiment" which consequently leads to the removal of a relationship between hidden parameter  $\theta$  and smoking  $S$ . This is to ensure only the causal connection of interest is observed. This seems to only satisfy the conclusiveness of the association, correlation or relationship between variables which is due to the causal link.

Without the controlled experiment, the results might be unsatisfactory. Thus it is more likely to discover the knowledge of causal relations rather than simply statistical associations that give some sense of genuine understanding, thus in such case, Bayesian networks provide a straightforward expression between variables [5].

In the extension of Wen, S et al. [12]'s state transition of Markov chains a stochastic technique whereby amongst others, the properties are state  $i$  and  $j$  communicate each is accessible from the other, and once it is in the state  $i$ , there is a positive probability that it will never return to state  $i$  and if the state is called an absorbing state if the probability of the state is absolute 1, say for instance  $P_{ii} = 1$ . Markov chain depends solely on the present state, not the prior or preceding states.

For each  $S_{ij}$ ,  $i$  represents the starting location and  $j$  represents the ending location for that move, where the row is the beginning location and the column is the ending location after one move. Each element in the matrix has the probability between 0 and 1, inclusive. The elements of each row of transition have the total probability size of 1 and the

matrix must be "squared" because it has row and column for each state. As previously stated, stochastic Markovian can usually be determined only by the current condition of the state in order to determine the outcome of the next state; it does not take into account the outcome of any of the previous states [11] and Bayesian network overcomes this limitation by taking consider of prior information through probabilistic inferential.

3) *Limitation of Previous Methods:* This section summarizes limitation of previous methods. Rules based are unbalanced or asymmetric in the sense that the statement of condition and the statement of consequence are not interchangeable (switchable), such that by observing the statement of consequence does not allow us to conclude the statement of condition [23], whereby the rule ignores some mechanisms or causal parameter of "causal direction" which makes this method is assured only up to certain level of precision and obviously, the rule based is inappropriate for representing the nature of causal relations amongst events.

K-means iteration process, on the other hand, is based on the ratio between two points. This will further distance the plausible parameter information as at the first place is has been reduced then the iteration makes the plausible information distance even further which finally will draw wrong information. Correlation statistics method seems to only satisfy the conclusiveness of the association, correlation or relationship between variables which is due to the causal link, thus it is more likely to discover the knowledge of causal relations rather than simply statistical associations that give some sense of genuine understanding [18].

Finally, stochastic Markovian can usually be determined only by the current condition of the state in order to determine the outcome of the next state; it does not take into account the outcome of any of the previous states [13].

### C. Bayesian Method Solutions

Having realized that rule-based, k-means, correlation and state-transition methods have limitations as a method of reasoning and knowledge representation, researchers switched their devotion towards a more sophisticated probabilistic interpretation of the certainty leading them towards the definition of Bayesian network approach [22].

Bayes's theorem had been pioneered by Tomas Bayes and published in a Posthumous Publication in 1763 which since then had been widely accepted as an uncontroversial result in probability theory [24]. This approach deals with decision-making process under uncertainty conditions or scenarios [25] which in summary it combines the past distribution (prior beliefs) with the available observed datasets to form the posterior distribution.

Bayesian is an inferencing tool that uses past observations (prior belief) to predict the future. Bayesian decision-making process provides an optimum result in classification problem when the prior or past probabilistic history is known which can further derive the estimation or expectation functions [24].

Bayesian Networks is also probabilistic causal networks also known as Belief Networks are the Artificial Intelligence framework for uncertainty supervision which is contrary to

deterministic approach to understand phenomena [26]. Although it was published in 1763 the techniques apply in health management and medicine decision-support systems are quite recent [24] and widely applied in clinical support decision [26]. Bayesian method offers instinctive, meaningful, professional and rational inferential analysis which gives the capability to solve complex situations given the priori distribution in addition to the dataset, thus making decisions easier to clarify and explain [25].

Previous Bayesian Quadratic and Bayesian Linear and models can mislead to false inadequate results due to the great size of parameters that have to be estimated from the dataset thus Naïve Bayes is capable a this issue [24]. Because of the intuitive ability to model uncertainty and complex chronological relationships amongst variables, Bayesian network is successfully applied in several research areas and domains [9,16] and the contribution of this paper is to tailor this general approach to generate new Bayesian Network detection technique to be applied at the Network-level environment [16].

1) *Bayesian Method in Clinical Expert Systems:* Expert systems development for clinical diagnosis has received a growing interest in the literature for the past few years [27]. Recent development of the expert systems particularly uses Bayesian is used in planning cardiac surgery for transfusion requirements [28].

Naranjo, L et al. [27] built a Clinical Expert System for the detection of PD using the Bayesian approach due to the traditional diagnosis which involves manual history taking is not definitive diagnostic test. Sometimes the procedure leads to misdiagnosis or even worst undiagnostic, thus there are great necessities to develop scientific systems which can help medical procedures especially in the neurological units. This research involved 80 subjects and 40 of them were healthy and another 40 infected by PD. The mean ( $\pm$  standard deviation) of age for the control and infected group was  $66.38 \pm 8.38$  and  $69.5 \pm 7.82$  respectively. Features selected were from 44 different acoustic sounds from five families of noise, amplitude, pitch, nonlinear and spectral.

Naranjo, L et al. [27] introduced  $n$  random variables  $Y_1 \dots Y_n$  which  $Y_i$  are observed is in Bernoulli distribution. The probability of success is  $P(Y_i = 1) = p_i$ ,  $i=1 \dots n$ . The probability of  $p_i$  are connected to two sets of covariates  $x_i$  and  $z_i$ , where  $x_i = (x_{i1} \dots x_{ik})^t$  is a  $K \times J$  matrix which is covariate  $K$  measured with  $J$  replicates, and  $z_i = (z_{i1} \dots z_{iH})^t$  is a  $H$  vector of a set of  $H$  covariates which are precisely identified. Then, suppose that  $x_{ij} = (x_{ij1} \dots x_{ijK})$  is the  $j$ th replication of the unknown covariates vector  $w_i = (w_{i1} \dots w_{iK})$  and assume their relationship is linear. This way  $x_{ij}$  are the substitutes or surrogates of  $w_i$ . The following model relates  $x_{ij}$  and  $z_i$ .

$$Y_i \sim \text{Bernoulli}(p_i),$$

$$\Psi^{-1}(p_i) = w_i^t \beta_x + z_i^t \beta_z \quad (3)$$

$$x_{ij} = w_i + \varepsilon_{ij} \quad (4)$$

$$\varepsilon_{ij} \sim \text{Normal}_K(0, G), \text{ indicates error vector}$$

where  $\beta = (\beta_x, \beta_z)^t$  is a  $(K + H)$  vector of unknown parameters, then  $\Psi^{-1}(\cdot)$  is a known nonnegative function or

link function ranges between 0 and 1, and  $G$  is a  $K \times K$  matrix of covariances and variances.  $\varepsilon_{ij}$  error vector is independent of  $w_i$ . Typically,  $\Psi(\cdot)$  is the CDF (cumulative distribution function) or normal distribution. To define prior distribution is to assumed normal distribution of the **regression** parameters,  $\beta \sim \text{Normal}_{K+H}(\mathbf{b}, \mathbf{B})$  and it is assumed prior distributions covariance and variance parameters  $G \sim \text{InvWishart}_K(\mathbf{V}, \nu)$ , where  $\mathbf{b}$ ,  $\mathbf{B}$ ,  $\nu$  and  $\mathbf{V}$  are fixed and  $w_i \sim \text{Normal}_K(\mu, \Sigma)$  where  $\mu, \Sigma$  are also fixed. Likelihood function is

$$l(\beta, G | y, x, z, w) = f(y | z, w, \beta) f(x | w, G) f(w), \quad (5)$$

then the posterior density is

$$\pi(\beta, G | y, x, z, w) \propto l(\beta, G | y, x, z, w) \pi(\beta) \pi(G) \quad (6)$$

where this approach makes use of the relationship between covariates and prior distributions to achieve posterior distributions. Fig. 6 below clearly shows the Bayesian model used in this research. Spotted easily several conditional probabilities networked together to the *priori* to derive *posteriori* distribution. Results are validated uses stratified cross-validation, but before results from precision (TP/TP+TP), recall rate (TP/TP+FN) and specificity (TN/TN+FP) are obtained.

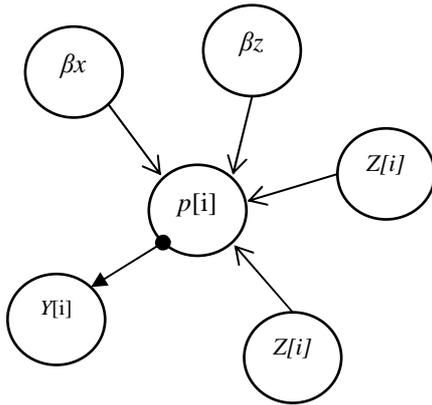


Fig. 6 A Bayesian model to determine Parkinson's disease

Next, Miasnikof, et al. [29] classified verbal autopsies (VA) which are largely adapted in low-income countries uses Naïve Bayes classifiers. This is due to no reasonable standard to validate the practice which had become the cause of home death, thus the studies pursue to measure results from Naïve Bayes classifier against existing standard procedures which is using physician classification. The dataset used is from Million Death Study, Matlab Bangladesh, and Agincourt South Africa. Sample sizes were 12,255 which contained deaths at ages 1 to 59 months, 15 to 64 years, 20 to 64 years and 28 days to 11 years. For every autopsy (VA), a probability (*priori*) will be assigned to each autopsy label, which is a specific feature in the form of symptom or sign, in accordance with their conditional probabilities. Only label with the maximum probability will be assigned to each label records. Suppose  $C_{j^*}$  is the cause of death, given a set of  $n$  records of signs and symptoms which are denoted by  $F_1 \dots F_n$ .

$$C_{j^*} = \arg \max_j \{Pr(C_j | F_1 \dots F_n)\} \quad (7)$$

Since the labels or features are chosen with the maximum probability, each feature  $F_n$  in the above equation is either 1 if the symptom is reported in VA and 0 if otherwise, and for simplification, the notation will be  $Pr(F_i = 1)$  or otherwise. Thus the proportional relationship of the above equation will be

$$Pr(C_j | F_1 \dots F_n) \propto Pr(C_j) \cdot Pr(F_1 \dots F_n | C_j) \quad (8)$$

this is the *posteriori*

Then is to apply the Bayesian assumption of the maximum probability to derive the following equation

$$Pr(C_j | F_1 \dots F_n) \propto Pr(C_j) \cdot \prod_{i=1}^n [j_i Pr(F_i | C_j) + (1-j_i) (1 - Pr(F_i | C_j))] \quad (9)$$

where,

$$j_i = \begin{cases} 1 & \text{if feature } i \text{ is reported} \\ 0 & \text{otherwise} \end{cases}$$

$$Pr(C_j) = |C_j| / N \quad (10)$$

where  $N$  is the size of features (total number of features)

$$Pr(F_i | C_j) = \frac{Pr(F_i \cap C_j)}{Pr(C_j)} = \frac{|F_i \cap C_j| / N}{|C_j| / N} = \frac{|F_i \cap C_j|}{|C_j|} \quad (11)$$

Priori distribution in this equation is just a proportion of sample cases  $|C_j|$ , against the total number of features. Whilst  $|F_i \cap C_j|$  is denoted as the total number of causes of death that showed feature  $F_i$ . The results will be evaluated against testing splits of the datasets to measure sensitivity or True Positive Rate (TPR) and specificity or False Positive Rate (FPR).

Meanwhile, Fuster-Parra, P et al. [26] applied Bayesian Network to determine the relationship between pertinent epidemiological signs of heart. For instance, given a set random variables  $X = (X_1 \dots X_n)$ . The prior distribution as a product of several conditional distributions is

$$P(X_1 \dots X_n) = \prod_{i=1}^n P(X_i | Pa(X_i^G)) \quad (12)$$

Where  $Pa(X_i^G)$  denotes the parent prior distribution which uniquely signifies the multivariate formulation in Bayesian Network which is also known as Bayesian Network chain rules. To enable inferential analysis, it is important to understand the flow of influence when any new information is introduced in Bayesian Network. For instance, let two variables  $X$  and  $Y$  which are separated by  $Z$  at several possible paths.  $X \rightarrow Z \rightarrow Y$  or  $X \leftarrow Z \leftarrow Y$  which is known as a serial connection,  $X \leftarrow Z \rightarrow Y$  as diverging connection and  $Z$  is initiated and finally,  $X \rightarrow Z \leftarrow Y$  which is known as converging connection where  $Z$  hasn't receive evidence. In this research, Bayesian was validated through 10-fold cross-validation uses log-likelihood loss function.

In medicine, Bayesian Network could characterize the conditional probabilistic value between symptoms and the diseases. Barbini, E et al. [24] shows a good illustration of

Bayesian Network applied in Clinical Expert Systems, where Bayesian Network is presented in a directed acyclic graph which has nodes and arcs. A node represents a random variable and an arc represents the conditional probability between nodes or variables.

Fig. 7 below illustrates their Bayesian Network model. For instance, between node A and node C, there is an arc representing the conditional probabilistic relationships. It also indicates that A has influence on C or A is the parent of node C as mentioned by [26]. There exist nodes, C and D, which are not connected to each other or lack of arc between them; this is to indicate that their existences are mutually exclusive and conditionally independence of each other. Nodes C and D which have parents' nodes are regarded by their conditional probabilities characteristic which are table systematically. Meanwhile, parents' nodes which in this case node A and B are regarded as *priori* or prior probability. Consequently,  $P(A)$  is denoted as the probability of any event A and  $x_{A'B}$  denotes the probability of any event C gave event B but not event A. On the occasion of all possible events of Bayesian Network have been defined, prior distribution of the parents' nodes and conditional probabilistic value of the inheritance should be specified.

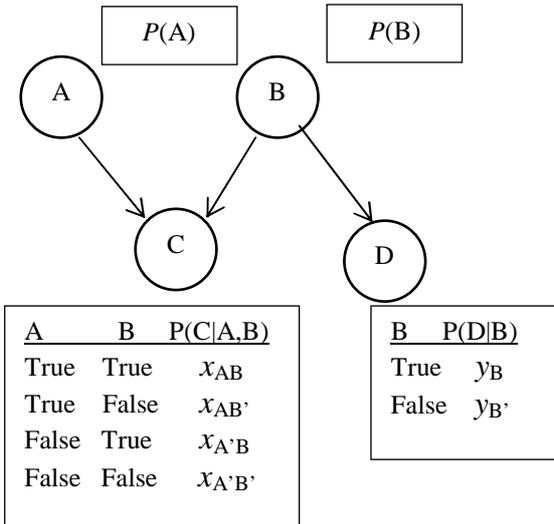


Fig. 7 A Bayesian model with two parent's nodes and two variables

As a conclusion, Bayesian Network is about several experts' structures which are combined or connected together to form a dependencies domain.

2) *Bayesian Method in Pattern Recognition*: Wang, et al. [30] proposed two and three nodes Bayesian Network model to recognise spatial age-related expression labels as shown in Fig. 8. Bayesian Network in this work represents a joint probability distribution amongst a set of features. As in Fig. 8, each node represents the features points and the arc or the link between nodes represents their conditional distribution probability which shows the probabilistic relationships between features points. Diverse Bayesian Network expressions were constructed to condition various spatial facial patterns amongst different expression and age.

This work was precluded with several activities such as forming two statistical hypothesis tests to scrutinise the

effect of age on spatial patterns of facial expression using ANOVA. After both hypothetical tests were conformed the two Bayesian Network model were applied. Bayesian Network discovers relationships amongst facial landmark points.

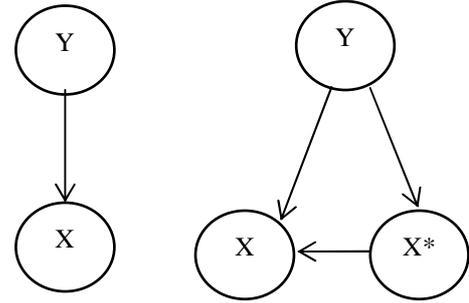


Fig. 8 Two nodes and three nodes Bayesian Network model

Throughout training a probabilistic model of  $P(x,x^*,y)$  is used, whereby the training set of  $(x_i,x^*_i,y_i)$  given  $i=1...l$  and  $x_i$  are features data in geometric distribution,  $x^*_i$  is the age data and  $y_i$  is the label of expression. The label has *priori* of  $P(y = k)$  ( $k = 1, 2, \dots, m$ ), where  $m$  is the size of expressions and the conditional probability of  $P(x|y = k)$  and  $P(x^*|y = k, x^*_i)$  are assessed within Maximum Likelihood method. It is understood that  $(x_i,x^*_i,y_i)$  is the training set where  $i=1...l$  and  $l$  is the size of training samples. The posterior probability of  $P(y = k|x)$  is calculated during this data training exercise and follow the following expression:

$$\begin{aligned}
 y^* &= \operatorname{argmax}_k P(y = k|x) & (13) \\
 &= \operatorname{argmax}_k \sum_{x^*} \frac{P(y = k, x, x^*)}{P(x)} \\
 &= \operatorname{argmax}_k \frac{P(y = k) \sum_{x^*} P(x^*|y = k) P(x | x^*, y = k)}{P(x)}
 \end{aligned}$$

Where conditional probability of  $P(x^*|y=k)$  can be characterised as Gaussian or Normal distribution  $P(x | x^*, y = k) \sim \text{Normal}(x | \mu_i^{(k)}, \Sigma_i^{(k)})$  where  $i=1,2,\dots,n$  and  $n$  is the size of the age groups for each  $x^*$  and  $x^*$  has  $n$  states. In this work, it is obvious that every  $x^*$  is converted or encoded into the conditional probability of  $P(y|x)$ .

Meanwhile for the probability of  $P(x|y)$  of the two nodes Bayesian Network model as in Fig. 3 also often conforms to a single Normal Distribution, however, the three nodes model with discrete  $x^*$  can improve the conditional distribution of  $P(x|y)$ . This research further constructed different Bayesian Network under different condition of expression and age. For instance, age group data is also regarded as controlled information, hence  $m \times n$  Bayesian Network model of  $G_c$  with  $c = 1...m \times n$  are constructed during the training period. Then, for every  $G_c$  parameter that is learned from the training set  $x_c = (x_{ci})_{i=1}^c$  where  $x_{ci} = (f_{ci}^1, f_{ci}^2, \dots, f_{ci}^p)$  and  $p$  is size of the features. This model is to learn the highest network score or the best  $x_c$ .

Suppose priori of  $G_c$  with  $c = 1...m \times n$  is uniform in distribution, thus during training we get  $P(G_c|x) \propto P(x/G_c)$ . Subsequently for every continuous node the probability are normal or Gaussian where the parameter is defined as  $f_j \sim \text{Normal}(b_j + W_j^T Pa(f_j), \delta_j^2)$ , where  $j = 1...p$ ,  $Pa(f_j)$  is the

parent's state of  $f_j$ ,  $W_j$  signifies regression coefficients,  $b_j$  denotes regression intercept and  $\delta_j^2$  shows the variance. Score function below shows the search strategy to learn  $G_c$ .

$$\begin{aligned} \text{Score}(G_c) &= \log P(x | G_c) \\ &= \max \log P(x | G_c, \theta_c) \end{aligned} \quad (14)$$

where  $\theta_c$  is the parameter given  $G_c$

The following Maximum Likelihood method is used to estimate the parameter of given  $G_c$  mentioned above, where  $\theta_c$  signifies parameter set of  $c$ th Bayesian model.

$$\theta_c = \operatorname{argmax} \log P(x | \theta_c) \quad (15)$$

Then the following expression is to signify the testing set into maximum likelihood method

$$c^* = \operatorname{arg max} c \in [1, m \times n] \frac{P(E_T | G_c)}{\text{Complexity}(G_c)} \quad (16)$$

Where  $E_T$  denotes sample features,  $G_c$  signifies the  $c$ th Bayesian model where the  $c$  ranges from 1 to  $m \times n$ ,  $P(E_T | G_c)$  denotes the probability of the sample features the  $c$ th model, and complexity signifies the complexity of  $G_c$  this is due to diverse differences amongst spatial structures, thus the probability of  $P(E_T | G_c)$  will be divided by complexity to seek the balance. Finally, the method is validated through ten-fold cross validation.

Meanwhile, Mihoub et al. [31] model face-to-face multimodal behavioural of co-verbal communication uses dynamic Bayesian Network method where it is a classical basic sophisticated multimodal bidirectional co-verbal communication which allows the partners to recurrently perceive, convey co-verbal movements such as body, hand and arm gestures and head movement.

Thirty games were constructed in which the trainer acted together with 3 different subjects or partners. The game objective is to place 10 cubes at random arrangement for which each game has a mean duration of 1 minute and 20 seconds. This interaction were modelled with 5 variables which are the IU denotes the interaction units which have another 6 different IUs of get, seek, point, indicate, verify and validate, the MP signifies manipulator gestures which have 5 values of rest, grasp, manipulate, end and none, the SP for the instructor speech with 5 values of cube, preposition, reference cube, else and none, the GT denotes the region of interest pointed out by the trainer's index finger which consists of 5 values of rest, target location, cube and reference cube and final variable is FX denotes gaze fixations of the trainer which have another 8 areas. FX and MP were annotated uses Pertech video. SP is transcribed uses speech recognition software. GT was annotated uses Qualys signals and finally IU is manually annotated on every gaze event. Elan software is used to manage multimodal scores.

Bayesian behavioural statistical model of Murphy [32] was used. Bayesian is considered as a probabilistic graphical model that provides conditional relationship representation of several stochastic conditions [33]. In Bayesian Network

an acyclic directed graphs of nodes and edges are represented uses conditional probabilities. For instance, an edge connecting between parent's node  $X$  to child's node  $Y$  signifies that node  $X$  has influence over node  $Y$ . Sometimes this relationship is learned from the data as in this research whereby the intra-slice structure is derived from K2 and REVEAL algorithm, whereby for K2 algorithm each node initially has no parents, thus this algorithm gradually, adds a parent. The order adopted was IU, MP, SP, GT and FX which shows the interaction unit stays in the first level and the sensory motor at the lower level as shown in Fig. 9.

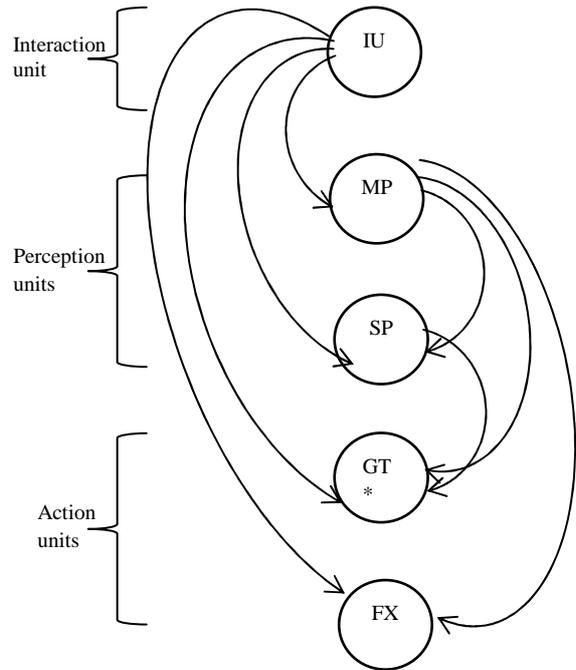


Fig. 9 The learned structure of Bayesian Network

Interesting properties have been discovered from this learned structure. The interaction unit, IU, stimulates both perception and action units. The MP impacts the SP, GT and FX. The SP or speech activity influences verbal or action behaviours of GT and FX. In essence, each random variable is influenced by its history or learned parents. For validation, this model is compared with the state-of-the-art baseline model which is Hidden Markov Model (HMM).

3) *Bayesian Method in Malware Analysis*: Kao, et al. [34] proposed a Bayesian of nonparametric approach to determine a malicious or benign at system level program under several first-order assumptions which have been modeled uses Dirichlet process mixture model. Let  $N$  represents the size of sample programs  $s$ , where  $s=1 \dots N$ . Dynamic trace of  $s$ th program is signified by  $\{Y_{s1} \dots Y_{sNs}\}$  where  $Y_{st} \in \{1 \dots M\}$  is the  $t$ th element during instruction call,  $N_s$  is the size of dynamic trace and  $M$  is the size of classes of instructions which implies that any similar instruction will be classed into similar group. To model first-order Markov structure it is enough to model distribution of first instruction  $Y_{s1}$  and  $M \times M$  transition matrix  $Z_s$

$Z_{src} = \sum_{t=2}^{N_s} I(Y_{st}=r, Y_{st}=c)$ , where  $(r,c)$  is the element of  $\{1 \dots M\}$

Both the first instruction and transition matrix distribution are permitted to fluctuate by programs. The first instruction of program  $s$  has probability distribution function of  $P(Y_{s1}=j)=q_{sj}$  with  $\sum_{j=1}^M q_{sj}=1$ . The subsequent or second instructions,  $t=2 \dots N_s$  are modelled as  $P(Y_{st}=c | Y_{st-1}=r)=P_{src}$  with  $\sum_{c=1}^M P_{src}=1$  for all  $r$ .  $P_s$  signifies the matrix of  $M \times M$  with  $(r,c)$ . Usually,  $N_s$  is very large that makes the first instruction  $Y_{s1}$ , is less information about the dynamic trace characteristic, thus it can be disregarded. In this case, matrix transition  $Z_s$  is sufficient for the dynamic trace statistic information, which then derives the following likelihood function for one single program of  $s$ ,

$$f(Z_s | P_s) = \prod_{r=1}^M \prod_{c=1}^M P_{src}^{Z_{src}} \quad (17)$$

The probability of the transition matrix  $P_s$  for a single program  $s$  is assumed  $P_s \sim F$  for any  $s$ , and where  $F$  probability distribution of matrices across malicious programs. The prior information of  $F$  is placed in non-parametric form to increase flexibility whereby  $F$  is assumed by Dirichlet process of Bayesian model

$$\begin{aligned} P_s &\sim F \\ F &\sim DP(a, F_b) \end{aligned} \quad (18)$$

Where  $a>0$  is concentration variable and  $F_b$  is the base distribution of the prior Dirichlet process. Large  $a$  pushes  $F$  *priori* closer to  $F_b$ , in this research  $F_b$  is distributed by matrix Dirichlet (MD),  $F_b = MD(yP)$ . Meanwhile, Dirichlet *priori* can be denoted as density mixture of

$$f(P_s | w, \lambda) = \sum_{k=1}^{\infty} w_k \cdot \delta(P_s | \lambda_k) \quad (19)$$

where  $\lambda = \{\lambda_k : k=1 \dots \infty\}$ ,  $w_k = \{w_k : k=1 \dots \infty\}$  and  $w_k > 0$  fulfil  $\sum_{k=1}^{\infty} w_k=1$ ,  $\delta$  signifies Dirac Delta density at point mass  $P_s = \lambda_k$  and  $\lambda_k \sim F_b$ .

However, this Dirichlet process has a discrete distribution which gives the transition probability is not appropriate, because the two probabilistic transitions of two executables could be identical and this is unrealistic. Therefore, the *priori* should be considered as in the next form of equation

$$\begin{aligned} P_s | \theta_s, \omega &\sim F = MD(\omega \theta_s) \\ \theta_s | G &\sim G \\ G &\sim DP(a, F_b) \end{aligned} \quad (20)$$

where the parameter  $\omega>0$  controls the variance of  $P_s$  distribution. By having this equation, probability of two identical transition matrices will be equal to 0. On the full model of Bayesian classification another variable is introduced,  $\xi_s$  where

$$\begin{aligned} \xi_s &= 1 \text{ if program } s \text{ is malicious} \\ &= 0 \text{ if program } s \text{ is benign} \end{aligned}$$

Where  $P(\xi_s=1)=\psi$  is the *priori* of being malicious. The following illustrates Bayesian model on a single program  $s$

$$\begin{aligned} f(Z_s | \xi_s = i, O_i) &= \sum_{k=1}^K w_k \cdot f_{MMP}(Z_s | \lambda_{ik}, o_i), i = 0, 1 \\ \lambda_{ik} &\sim MD(y_i P_i) \\ \xi_s &\sim \text{Bern}(\psi) \end{aligned} \quad (21)$$

where  $O_0 = \{\lambda_0, y_0, o_0, w_0, K_0\}$  and  $O_1 = \{\lambda_1, y_1, o_1, w_1, K_1\}$  are all the pool of parameters beneath the malicious and benign models that allow different malicious and benign programs that are with different mixture densities and base distributions. Finally, the validation is against existing support vector machine (SVM) and elastic net logistic (ENL) regression method. In the case of generating a parametric form of the conditional probability functions is not possible Non-parametric approach is used, however, this approach is complex [1] and will not be suggested as a future research. However, the essence of establishing prior, conditional and posterior distribution as in the Bayesian theorem will be applied in the proposed method.

On the other hand, Weaver [35] modeled bot net scanning behaviour in a large network environment as she claimed the sandbox environments seldom emulate real user experience. Furthermore, NAT and DHCP configuration have made the IP in ISP level doesn't map one-to-one and the real perpetrator will not be revealed. However, from the point of view of experienced network engineer or administrator, it is true that the NAT will translate outbound and inbound traffic, but each organisation will be assigned unique IP address to make them online in the large network, however it is the work of local administrator to stream down or deep inspect the flow to identify the source and destination IP within the local area network.

The monitored network was from large private network from the period approximately 2 months window from the month of March 5<sup>th</sup> until the month of April 24<sup>th</sup>. The network flows were collected uses flow analysis toolset of SiLK or System for Internet-Level Knowledge which is available at [tools.netsa.cert.org/silk](http://tools.netsa.cert.org/silk) and the features monitored were from the source IP address, timestamp and source port. Accumulated around 33.6 million total unique IP addresses were monitored disbursed across 1.1 million /27 IP blocks. Behaviours show fluctuated events of periodic inactivity which indicates consistent machine shutting down activities and several spikes which indicate starting of new activities. The model was from the single machine connection rates of several user activities such as web surfing or switching the machine off or on. This is the alternative behavioural model of counting infected machines from the layered network blocks or IP addresses.  $\lambda_s$  denotes the Poisson mean of the number of connection requests,  $y_t$  in,  $t$  hour ( $y_t \sim \text{Poisson}(\lambda_t)$ ), where  $\lambda_t = q(1 + a^k(\omega - 1))$  such that  $q$  denotes the baseline rate,  $a$  denotes the decay rate,  $k$  denotes a spike after an active hour and  $\omega$  denotes a spike multiplier. A steady rate is denoted by  $q$  (baseline rate) and occurs when the host is idle. The state of the user's machine or system is denoted by  $\eta$ , represented by three states of possibilities which are "o=off, s=spike, d=decay". A host is said to be in the "off" state when there is no connection request,  $y_t = 0$  with probability=1.0. When a host is active, three conditions are applicable; a geometric decay at rate  $a$ , a multiplier spike at  $\omega > 1$  and a baseline at rate  $q$ .

A set of transition probabilities between spike and decay of the user's state  $\eta$  to  $\eta+1$  hour-to-hour is modelled uses

geometric distribution with rates of  $y_i$  and  $y_d$ , meanwhile transitions between non-equal user's state, off-to-spike, spike-to-decay and off-to-decay is modelled uses sine wave with maximum height of  $p$  and scaling amplitude of  $v$  within 24-hour cycle and  $t^*$ , an hour a day will be between 0 to 23. Thus the transitioning state between  $s_2$  to  $s_1$  is denoted by

$$P_{s_2 | s_1}(p_{s_1}, v_{s_1}, t^*) = (p_{s_1} / (v_{s_1} + 2)) [\sin(2\pi t^* / 24) + v_{s_1} + 1] \quad (21)$$

Next, is to build probability model of machine and user's activity uses Gibbs sampling a method of Markov Chain Monte Carlo in order to explore the posterior distributions [17] by considering several relevant prior information. Suppose  $\eta$  is the machine state at time  $t$ :  $\eta_t \in \{o, s, d\}$  and  $\eta = \{\eta_0 \dots \eta_T\}$  and  $y = \{y_0 \dots y_T\}$  as the observed vectors, subsequently let  $\psi = \{p_{\{o, s, d\}}, v_{\{o, s, d\}}, y_{\{o, s, d\}}, \lambda_{off}\}$  be the probability of user's transitional activities and  $\theta = (g, w, a)$  be the machine-level parameters. The likelihood would be the joint of all distribution given earlier,  $l(\psi, \theta, \eta; y)$ .

Gibbs sampling explores *posteriori* by initiating value for the entire parameters and iteratively straws new values for a subsequent subset based on the conditional probability of those parameters and Markov chain derived from this chains of iteration is the joint *posteriori* for the entire parameters [8]. For instance, let  $\varphi \in (\psi, \theta, \eta)$  with *priori* of  $\pi(\varphi)$  and let  $\{\psi, \theta, \eta\} | \varphi$  be the observed vectors for the entire parameters except for  $\varphi$ . The complete general conditional probability distribution is

$$\pi(\varphi | \{\psi, \theta, \eta\}_{|\varphi}, y) = \frac{\pi(\varphi) l(\varphi, \{\psi, \theta, \eta\}_{|\varphi}; y)}{\int_{\varphi} \pi(\varphi) l(\varphi, \{\psi, \theta, \eta\}_{|\varphi}; y)} \quad (22)$$

Boukhtouta, et al. [36] compared several machine learning techniques such as J48, Naïve Bayesian and Support Vector Machine to detect malicious activity at the network level as the state-of-the-art Intrusion Detection Systems (IDS) uses signature-based techniques to filter bad traffic is insufficient. The non-malicious traffic was obtained from Defence Advanced Research Project Agency (DARPA) trusted source. However, the research is a software suite-based approach which makes it difficult to evaluate the analysis engine.

### III. RESULTS AND DISCUSSION

The previous section discusses various application of Bayesian Network in several domains like Clinical Expert Systems, Pattern Recognition and reminiscent of Bayesian work in malware analysis. In essence, Bayesian network is a directed acyclic probabilistic model that represents a set of random variables and their conditional probabilistic relationships.

There are a few properties encompasses Bayesian Network method. Statistically, a Bayesian Network model has four properties which are 1) prior probability or *priori*, 2) the likelihood or the conditional probability, 3) posterior probability or *posteriori* and finally 4) the relationship of parents' nodes and its inheritance. For instance, given a prior information of  $y$  and parameter  $\theta$ , the *priori* would be  $p(\theta)$  and the conditional probability is  $p(y | \theta)$  and generated posterior would be

$$p(\theta | y) \propto p(y | \theta). p(\theta) \quad (23)$$

Frequently in Bayesian Network, the *priori* rest on to other parameters  $\varphi$  that are not declared in the conditional probability or the likelihood, thus the prior  $p(\theta)$  must be substituted by the conditional probability of  $p(\theta | \varphi)$ . The newly elected parameters  $\varphi$  *posteriori* would be

$$p(\theta, \varphi | y) \propto p(y | \theta). p(\theta | \varphi). p(\varphi) \quad (24)$$

Based on the literature, studies from Clinical Expert Systems have the full four properties in their expert systems model, whilst domain like pattern recognition and existing malware analysis studies fulfill the first three out of four of Bayesian or Naïve Bayes properties. The four properties discussed here addresses problems in the common criteria.

### IV. CONCLUSIONS

This paper discusses the application of Bayesian Network model in various domains such as Clinical Expert Systems, Artificial Intelligence, Pattern Recognition and reveals any potential approach available in the domain of Computer Networks. It is discovered that Bayesian (Naïve Bayes) method has been applied in the malware analysis domain both in network and system-level behavioural analysis. However, this method has several issues in dealing with zero-day attacks which are related to the issue of predicting future attacks pattern. Based on the literature Bayesian Network properties which have been applied in various domains could have the potential to overcome those problems. Thus these properties could be used as guidance for future studies on modeling Behavioural Malware Predictive Analytics at the network level.

### ACKNOWLEDGMENT

We would like to thank Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia for supporting this research using research grant code DPP-2014-014.

### REFERENCES

- [1] M. H. M. Yusof and M. R. Mokhtar, "The Nomenclature of Behavioural Malware Analysis Methods in Computer Networks". *Security and Communication Networks*. Id: SCN-16-0424, 2016
- [2] D. S. K. Rai. "Performance Based Comparative Analysis of AODV, DSR and DSDV Protocols". Master Thesis. Computer Science and Engineering Department Thapar University. 2009.
- [3] H. D. Trung, W. Benjapolakul, and P. M. Duc, "Performance evaluation and comparison of different ad hoc routing protocols," *Journal of Computer Communications*, vol.30, pp. 2478-2496, 2007.
- [4] P. D Hoff, *A First Course in Bayesian Statistical Methods*, Seattle WA, USA: Springer. 2009.
- [5] T. Koski, and J. M. Noble, *Bayesian Networks*, United Kingdom, John Wiley & Sons, Ltd. 2009.
- [6] B. Rahbarinia, R. Perdisci, and M. Antonakakis, "Segugio: Efficient Behavior-Based Tracking of Malware-Control Domains in Large ISP Networks", in *45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, 2015, pp. 403-414.
- [7] M. Zaman, T. Siddiqui, M. R. Amin, and M. S. Hossain, "Malware Detection in Android by Network Traffic Analysis," *Next Generation Mobile Apps, Services and Technologies (NGMAST)*, pp. 66-71, 2015
- [8] E.I. Edem, C. Benzaid, A. Al-Nemrat, and P. Watters, "Analysis of Malware Behaviour: Using Data Mining Clustering Techniques to Support Forensics Investigation", in *Fifth Cybercrime and Trustworthy Computing Conference*, 2014, pp. 54-63.

- [9] A. Feizollah, N. B. Anuar, R. Salleh, and F. Amalina. "Comparative study of k-means and mini batch k-means clustering algorithms in android malware detection using network traffic analysis", in *International Symposium on Biometrics and Security Technologies (ISBAST)*, 2014, pp.193 - 197.
- [10] L.J.G. Villalba, A. L. S. Orozco, and J. M. Vidal,. "Malware Detection System by Payload Analysis of Network Traffic," *IEEE Latin America Transactions*,13(3), 2015.
- [11] P. O'Kane, S. Sezer, and K. McLaughlin. "N-gram density based malware detection," in *World Symposium on Computer Applications & Research (WSCAR)*, 2014, pp. 1-6.
- [12] S. Wen, W. Zhou, J. Zhang, Y. Xiang, W. Zhou, W. Jia, and C. C. Zou, "Modeling and Analysis on the Propagation Dynamics of Modern Email Malware". *IEEE Transactions on Dependable and Secure Computing*, 11(4), pp. 361 - 374.
- [13] J. G. Kemeny, J. L. Snell, A. Knapp, *Markov Chains*, Pearson Education, Inc. 2003.
- [14] M. A. Ahmad and S. Woodhead, "Containment of Fast Scanning Computer Network Worms," in *Proceedings of the 8th International Conference on Internet and Distributed Computing Systems*, 2015 vol 9258, pp. 235-247.
- [15] L. Xue, and G. Sun, "Design and Implementation of Malware Detection System based on Network Behavior," *Security And Communication Networks*, Volume 8 (3), pp. 459-470, Feb. 2015.
- [16] A. Arora, S. Garg, and S. K. Peddoju, "Malware Detection Using Network Traffic Analysis in Android Based Mobile Devices," in *Eight International Conference on Next Generation Mobile Apps, Services and Technologies*, 2014, p. 66-71.
- [17] P. S. Mann, *Introductory Statistics*, 8<sup>th</sup> Edition. Wiley, ISBN: 978-1-118-17224-7, 2012.
- [18] D. Sorensen, and D. Gianola, *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. New York, USA: Springer, 2002.
- [19] K. Patel and B. Buddadev, *Detection and Mitigation of Android Malware Through Hybrid Approach*, J. H. Abawajy, S. Mukherjea, S. M. Thampi, A. Ruiz-Martinez, Ed., *Security in Computing and Communications: Third International Symposium, SSCC 2015*, Kochi, India, pp. 455-463.
- [20] U. B. Kjaerulff and A. L. Madsen. *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*. Aalborg, Denmark, Springer, 2008.
- [21] (2016) *Health and Social Care Information Centre. Bronchitis*. [Online]. Available: <http://www.nhs.uk/conditions/Bronchitis/Pages/Introduction.aspx> website 2016
- [22] R. E. Neapolitan. *Learning Bayesian Networks*. Northeastern Illinois University, Chicago, Pearson Prentice Hall, 2004.
- [23] H. D. Trung, W. Benjapolakul, and P. M. Duc. "Performance evaluation and comparison of different ad hoc routing protocols". *Journal of Computer Communications*, vol. 30, pp. 2478-2496, 2007.
- [24] E. Barbini, P. Manzi, and P. Barbini. *Bayesian Approach in Medicine and Health Management*, A. J. Rodriguez-Morales, Ed., 2013.
- [25] J. W. Sheppard, and M. A. Kaufman, "A Bayesian approach to diagnosis and prognosis using built-in test," *IEEE Transactions on Instrumentation and Measurement*, vol. 54(3), pp. 1003-1018, 2005
- [26] P. Fuster-Parra, P. Tauler, M. Bennasar-Veny, A. Lig, eza, A. A L'opez-Gonz'alez, A. Aguil'o. "Bayesian Network Modeling: a Case Study of an Epidemiologic System Analysis of Cardiovascular Risk". *Computer Methods and Programs in Biomedicine*, 2015.
- [27] L. Naranjo, C. J Perez, Y. Campos-Roca, and J. Martin. "Addressing voice recording replications for Parkinson's disease detection", *Expert Systems With Applications*, May 2016.
- [28] G. Cevenini, E. Barbini, M. R. Massai, and P. Barbini, "A naive Bayes classifier for planning transfusion requirements in heart surgery" *Journal of Evaluation in Clinical Practice*, vol.19(1), pp 25-29, 2013.
- [29] P. Miasnikof, V. Giannakeas, M. Gomes, L. Aleksandrowicz, A. Y. Shestopaloff, D. Alam, S. Tollman, A. Samarikhalaj, P. Jha, "Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths," *BMC Medicine*, vol. 13(1), 2015.
- [30] S. F. Wang, S. Wu, Z. Gao, Q. Ji, "Facial expression recognition through modeling age-related spatial patterns," *Multimedia Tools And Applications*, vol. 75(7), pp. 3937-3954, 2016.
- [31] A. Mihoub, G. Bailly, C. Wolf, F. Elisei, "Graphical models for social behavior modeling in face-to face interaction", *Pattern Recognition Letters*, vol. 74, pp. 82-89, 2016.
- [32] K. P. Murphy, "Dynamic Bayesian networks: representation, inference and learning", PhD Thesis, University of California, Berkeley, 2002.
- [33] D. Koller, and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.
- [34] B. Kao, C. Reich, Storlie, and B. Anderson, "Malware Detection Using Nonparametric Bayesian Clustering and Classification Techniques", *Technometrics*, vol. 57(4), pp. 535-546, 2015.
- [35] R. Weaver, "Visualizing and Modelling the Scanning Behavior of the Conficker Botnet in the Presence of User and Network Activity". *IEEE Transactions on Information Forensics And Security*, vol. 10(5), pp. 1039-1051, May 2015.
- [36] A. Boukhtouta, N. E. Lakhdari, S. A. Mokhov, and M. Debbabi, "Towards Fingerprinting Malicious Traffic," *Procedia Computer Science*, vol. 19, pp. 548-555, 2013.
- [37] M. H. M. Yusof, and M. R. Mokhtar, "A Review on Taxonomy of Malware Analysis Studies," *Advanced Science Letters*, vol. 22, 2016.
- [38] S. Nari, A. A Ghorbani, "Automated Malware Classification based on Network Behavior," in *International Conference on Computing, Networking and Communications, Communications and Information Security Symposium*, 2013, pp. 642-647.