IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

Advanced hybrid stochastic dynamic Bayesian network inference algorithm development in the context of the web applications test execution

To cite this article: T V Azarnova and P V Polukhin 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **537** 052028

Advanced hybrid stochastic dynamic Bayesian network inference algorithm development in the context of the web applications test execution

T V Azarnova and P V Polukhin

Department of Applied Mathematics and Informatics and Mechanics, Voronezh State University, Voronezh, Russia

E-mail: ivdas92@mail.ru

Abstract. The article is devoted to the application of dynamic Bayesian networks models for fuzzing web applications and development of effective hybrid algorithms for probabilistic inference based on particle filter algorithm. Dynamic Bayesian networks models allow to simulate the dynamic process transformation of web applications associated with the process of their constant instrumental and logical updates, and create a probabilistic structure required for learning process of testing the top web applications vulnerabilities, that able to use the evidence and inference results obtained in the retrospective and current testing slices and improve testing mechanisms in new time slices. The hybrid probabilistic inference algorithm for dynamic Bayesian networks models for testing web-applications, proposed in the current research, significantly increase the efficiency of the classical approximate probabilistic inference algorithms, well reflect the features of the temporary testing links formation and adapted to the detection of anomalous errors.

1. Introduction

At the present stage of information systems and technologies development, web-applications are widely used in various areas, such as education, health, business, banking, industrial production, etc. Development of web-applications are engaged in both professional teams of web-developers and teams that do not specialize only in web-applications development, that create applications for the custom needs of the such sphere in which the company operates. Web applications are based on the interaction between the browser and the web server, which creates important advantages for both developers and end users. The developed applications are cross-platform, not sensitive to the operating system, there are tools for continuous updating and expanding the capabilities of applications, there is no need to download programs and updates, because the applications are hosted on a centralized server available over the Internet. One of the existing shortcomings of web-applications today is the problem related to their information security. Modern statistical researches show that a high percentage of analyzed web-applications contain vulnerabilities of various types The existence of vulnerabilities is closely connected to the developers errors and uncoordinated development process, in particular the updates, different mechanisms for developing web applications. Aggregation components of information systems and technologies intended to design, build, upgrade, and configuration of the web applications interaction represent multicomponent, multiplatform, continuously evolving in accordance with trends projected requirements. The rates of various components renewal are not synchronized; there is no single mechanism for managing the platform development as a single system with systemic principles of interaction.

During the employment of an insecure application, a third-party user may have access to resources that are processed within the application. In this regard, there is a need to standardize approaches to the applications development, creation of various testing tools, as well as the development of a wide range of requirements that must be considered at the stage of software development and maintenance. The most interesting security approaches regulated by the next information security projects – OWASP and MITRE, which establish not only cause-and-effect relationships between different groups of errors inherent in a particular application, but also contribute to the formation of the regulatory framework necessary for the localization of different types of errors. Increasing the functionality and adaptability of application design to the needs of users leads to rise a class of significant errors associated with undocumented formation and sending requests in the context of a registered user on any information resource. Such errors are quite common and are called cross-site request forgery (CSRF). The whence of this error is due to insufficient filtering of input parameters received by the application and further participating in the formation of the program logic. Today there are a wide range of approaches related with creating effective technologies for testing vulnerabilities of web applications. One of the most advanced testing technologies is fuzzing, which allows implementing complex testing at various stages of the application life cycle. As part of the fuzzing technology development, the article examines the use of dynamic Bayesian networks for learning the testing process to perceive the evidence, obtained at different time slices of testing and in the terms of structured probabilistic inference information to predict the presence of certain vulnerabilities in the current or future testing slices. The introduction of these technologies can significantly improve the efficiency and resource efficiency of fuzzing testing process.

2. Hybrid probabilistic inference algorithms for dynamic Bayesian networks based on importance sampling approach

The adaptation of mathematical models for the analysis dynamic web applications testing is an important direction in the design and development of modern fuzzing tools. The most suitable to the solution of testing problems are dynamic Bayesian networks (DBN), which allow forming conditional probabilistic relations between the test data sets, generated for the analysis a specific types of errors and the conditions when the such types of errors occurrence. Let us consider briefly the main aspects of the DBN application for modeling the testing process and the solution of the main testing tasks in the language of probabilistic inference of Bayesian networks apparatus. DBN is a set of static Bayesian networks taken at the fixed time interval $(t;t+k)$. Each node of the Bayesian network is described by a table of conditional probabilities (CPT) formed in accordance with the criterion of conditional independence (CI). The conditional independence criterion for Bayesian networks be in that each node y for known values of parents Y is independent of any set X such that $x \notin Y$ and X Ъ Y . CI for each of the vertices is directly related to the notion of Markov Blanket (MB). MB is a set of vertices for a DBN node that includes the vertex itself, its parent vertices, and the parents of its child vertices. MP can be obtained directly from distribution $P(Y)$

$$P(Y)=\sum_{x_{r_1}}\sum_{x_{r_2}}\cdots\sum_{x_{r_{k-1}}}\prod_{x_i \notin child^*(x_{r_k})} P(x_i \mid Parent(x_i))\times\sum_{x_{r_k}}\prod_{x_i \in child^*(x_{r_k})} P(x_i|Parent(x_i)), x_{r_k} \in X \setminus Y, \quad (1)$$

All nodes of DBN structurally divided into two categories: the unobserved (hidden) variables X_t and the evidence variables E_t. DBN nodes that are connected to the nearby slices form the transition model $P(X_{t+1} \mid X_t)$. Nodes that are evidences for the moment t, form sensor model $P(E_t|X_t)$.

In this article, the technology of using dynamic Bayesian networks to the web applications testing process will be demonstrated on the example of CSRF testing. The topology of the dynamic Bayesian CSRF test network consisting of two time slices shown in figure 1

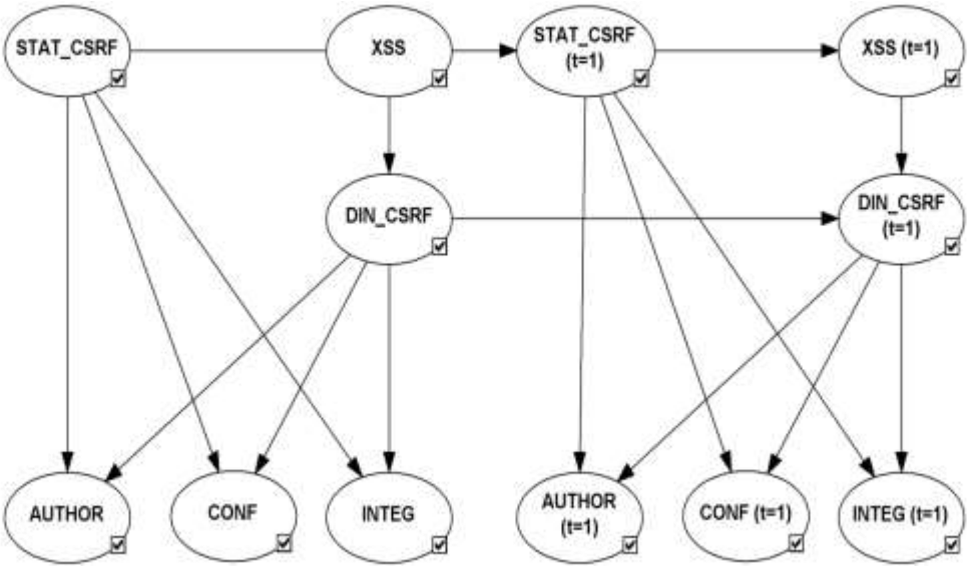

Figure 1. Dynamic Bayesian network of the CSRF errors testing process.

Table 1. Description of the nodes in CSRF DBN errors testing.

Node name	Description
XSS	The possibility of using XSS as a delivery transport
STAT_CSRF	The mechanism of static CSRF execution, through the simple simulated request to the target resource
DIN_CSRF	Mechanisms to dynamically generate and send CSRF requests
AUTHOR, CONF, INTEG	Violation of confidentiality, integrity and authorization mechanisms

To obtain a posteriori probabilistic distribution of dynamic Bayesian nodes for the time slice $t+k$, the technology of probabilistic inference based on the sequential Markov Chains Monte Carlo method is used. One of the algorithms that implement this policy is the partial filter algorithm using the method of importance sampling (PFIS). The algorithm is based on the method of calculating the probability distribution and the computation of the generated samples weights

$$\mathrm{P}\left(\mathrm{X}_{t+1}|\mathrm{E}_{1:t+1}\right)=\sum_{i=1}^{N_s}\mathrm{W}_i^{t+1}\delta\left(\mathrm{X}_{t+1},\mathrm{X}_{t+1}^i\right), \tag{2}$$

where $X_{(t+1)}^{i}$ is the sample corresponding to the variable X_{t+1}

The formation of samples according to the probability distribution $\xi(X)$ is quite difficult, it is possible to determine the importance distribution $q(X)$, in this case $\mathrm{W}_i=\xi(\mathrm{X})/\mathrm{q}(\mathrm{X})$. After computation of all weights $\mathrm{W}=\left(\mathrm{W}_1,\mathrm{W}_2,\ldots,\mathrm{W}_n\right)$ it is necessary to perform weights data normalization $\sum_{i=1}^{n}\mathrm{W}_i=1$. The given weights taking into account evidence E will be proportional to the following ratio:

$$W_i \propto \frac{P(X^i_{1:t+1} \mid E_{1:t+1})}{q(X^i_{1:t+1} \mid E_{1:t+1})} \tag{3}$$

To determine the probability distribution $P\left(X_{t+1}|E_{1:t+1}\right)$, it is necessary to offer the importance distribution density in the following factorized form

$$q\left(X_{1:t+1}|E_{1:t+1}\right) = q\left(X_{t+1}|X_{1:t}, E_{1:t+1}\right) q\left(X_{1:t}|E_{1:t}\right) = q\left(X_t\right) \prod_{t=2}^{n} q(X_{t+1} \mid X_{1:t}, E_{1:t}) \tag{4}$$

From the expression (4) follows that we can generate sample $X^i_{1:t+1} \sim q\left(X_{t+1}|E_{1:t+1}\right)$ multiplying each value of the sample $X^i_{1:t} \sim q\left(X_t|E_{1:t}\right)$ on the new state $X^i_{t+1} \sim q(X_{t+1} \mid X_{1:t}, E_{1:t+1})$. The distribution value $P\left(X_{t+1}|E_{1:t+1}\right)$ can be obtained by recursively using the Bayes rule

$$P\left(X_{t+1}|E_{1:t+1}\right) = \frac{P\left(E_{t+1}|X_{t+1}\right) P\left(X_{t+1}|X_t\right)}{P\left(E_{t+1}|E_{1:t}\right)} \times P\left(X_{1:t}|E_{1:t}\right), \tag{5}$$

$$P\left(X_{t+1}|E_{1:t+1}\right) \propto P\left(E_{t+1}|X_{t+1}\right) P\left(X_{t+1}|X_t\right) P\left(X_{1:t}|E_{1:t}\right) \tag{6}$$

Inserting expressions (5) and (6) into formula (3) we obtain the desired weight distribution of W_i for all generated samples

$$W_i \propto \frac{P\left(E_{t+1}|X^i_{t+1}\right) P\left(X^i_{t+1}|X^i_t\right) P\left(X^i_{1:t}|E_{1:t}\right)}{q\left(X^i_{t+1}|X^i_{1:t}, E_{t+1}\right) q\left(X^i_{1:t}|E_{1:t}\right)} = W^t_i \frac{P(E_{t+1} \mid X^i_{t+1}) P(X^i_{t+1} \mid X^i_t)}{q(X^i_{t+1} \mid X^i_{1:t}, E_{1:t+1})} \tag{7}$$

The formula (7) can be used to solve the filtration problem (probability distribution $P(X_{t+1} \mid E_{1;t+1}))$, $q\left(X_{t+1}|X_{1:t}, E_{1:t+1}\right) = q(X_{t+1} \mid X_t, E_{t+1})$. The density of the importance sample distribution depends only on X_t and E_{t+1}, and as a consequence it is necessary to keep only the values X^i_{t+1}, while the chain of events X^i_t and evidence $E_{1:t+1}$ can be excluded. The expressions for calculating the updated weights W^{t+1}_i take the following form

$$W^{t+1}_i \propto W^t_i \frac{P(E_{t+1} \mid X^i_{t+1}) P(X^i_{t+1} \mid X^i_t)}{q(X^i_{t+1} \mid X^i_t, E_{1:t+1})} \tag{8}$$

In the process of generating importance weights, it is often assumed that the formation of the importance distribution for each generated samples performed from a priori distribution

$$q\left(X_{t+1}|X^i_t, E_{t+1}\right) = P(X_{t+1} \mid X^i_t) \tag{9}$$

Then, around the expression (9), expression (8) can be reduced to the following simplified form

$$W_i = W^t_i P(E_{t+1} \mid X^i_{t+1}) \tag{10}$$

Based on the formula (10) standard importance samples algorithm may be converted to a sequential importance samples algorithm (SIS). In this case, the number of N_{eff} samples can be obtained from the following expression

$$N = \frac{N_s}{1 + E\left(\omega^{t+1}_i\right)}, \; N \le N_s, \tag{11}$$

where $E\left(\omega_i^{t+1}\right)=P\left(X_{t+1}^i|E_{1:t+1}\right)/q(X_{t+1}^i|X_t^i,E_{t+1})$.

The expression (11) cannot be calculated exactly, but the corresponding value $\hat{N}$ can be determined as

$$\hat{N}=\frac{1}{\sum_{i=1}^{N_s}\left(W_i^{t+1}\right)^2} \tag{12}$$

where W_i^{t+1} are normalized weights obtained from the expression (8).

Meanwhile, the algorithm has one drawback; its time complexity is proportional to the number of samples that necessary generates to achieve the required level of algorithm accuracy. It is proposed to use a hybrid algorithm by combining algorithms of PF and metropolis-Hastings (MH). MH is used in the process of generating samples from the distribution $Q(X,X')$ converging to the distribution $P(X)$. Applying the approach embodied in the MH algorithm, the probability of transition from the X in X' state can be written as the following equation

$$\varphi(X,X')=\min\left[1,\frac{P(E|X')P(X')Q\left(X|X^{'}\right)}{P(E|X)P(X)Q(X'|X)}\right], \tag{13}$$

Assuming that the distribution of $Q\left(X|X^{'}\right)$ is symmetric, the expression (13) can be reduced to the following form

$$\varphi(X,X')=\min\left[1,\frac{P(E|X')}{P(E|X)}\right], \tag{14}$$

In this case, we can define an expression for the Markov chain corresponding to the transition probability $P(X'|X)$

$$MC=P(X'|X)=Q(X')\times\min\left[1,\frac{W(X')}{W(X)}\right], \tag{15}$$

where $W(X)=P(X)/Q(X)$ is the importance weights distribution.

In the experimental part of the scientific research, a simulation of the test environment consisting of ten applications with the possibility of protection components optional inclusion. This allows simulating release of software updates and provides an opportunity to represent testing as a temporary model build upon the DBN. As part of the test environment modeling, we used applications developed and maintained by the OWASP consortium and deployed on two types of web servers: Apache and Nginx. As a result of the initial testing, the accumulation of statistical data was made, around which the DBN for the CSRF group of errors was built, determined a priori distribution $P(X_0)$, transition $P(X_{t+1}|X_t)$ and perception $P\left(X_{t+1}|E_{t+1}\right)$ models.

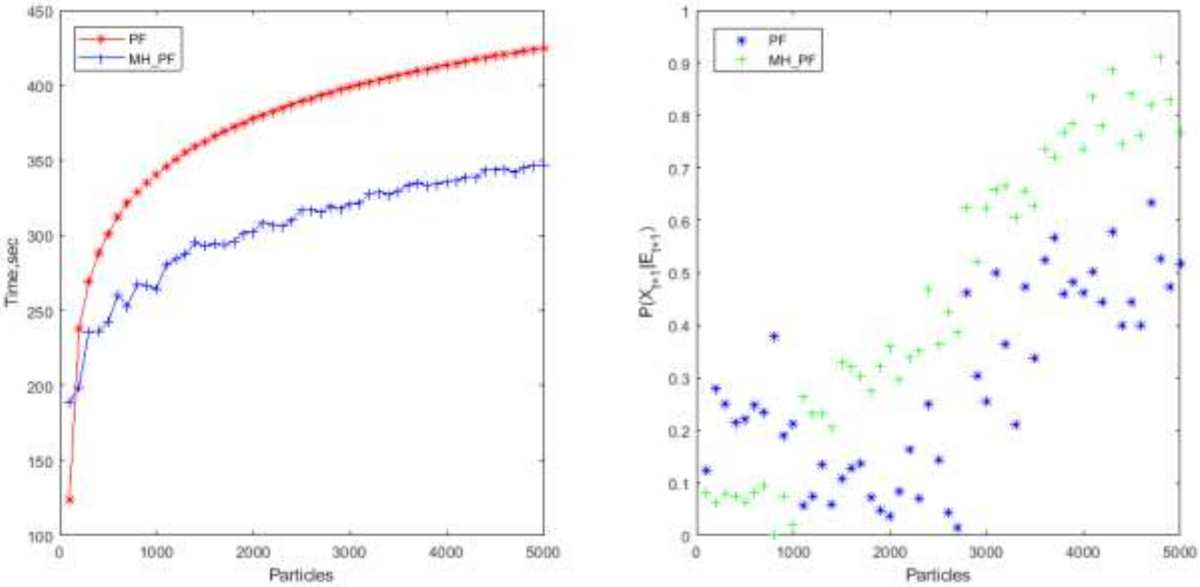

Figure 2. Probability distributions and resource efficiency of algorithms PF and algorithm PF with the use of MH.

Analysis of figure 2 shows that the hybrid probabilistic algorithm PF with the use of MH employment allows obtaining high computational performance characteristics and improving the efficiency of the classical algorithm PA. PF with MH algorithm allows reducing the cost of subsequent samples generation and increases the efficiency in relation to the generic PF algorithm, as after iterative regeneration of samples using MH we can obtain a high quality estimates properties for each of the DBN variables. The MH employment helps to optimize the sampling procedure and obtain as the final result the probability distribution right up to the state $t+k$ hidden variables, on the assumption there is evidence obtained for the sequence of time slices located before the time slice $t+k$. Adaptation of the DBN and the investigated hybrid stochastic inference algorithm allows obtaining the necessary set of test rules that detecting CSRF errors group with a sufficiently high probability level.

3. Conclusion

Dynamic Bayesian networks are a modern, but at the same time well-tested tool for stochastic dynamic processes analyses. Adaptation of dynamic Bayesian networks for testing methods and fuzzing mechanisms implementation, allows optimizing the procedure of test data generation, to adjust the internal testing algorithm mechanisms via solving probabilistic inference problems. The hybrid algorithms application allows to predict the possible state of the application and to develop recommendations, necessary for the up to date configuration in test data. The use of the proposed tools makes it possible to optimize the errors search procedure, simplify the test data generating and generally improve the efficiency of the web applications testing process. The proposed tools are particularly relevant for companies engaged in security audit, as well as companies specializing in software development in various industries. The proposed tools are particularly relevant for companies engaged in security audit, as well as companies specializing in software development in various industries.

References

[1] Kelbert M and Suhov U 2007 *Probability and statistics by example. Basic Probability and Statistic* (Cambrige: Cambridge university press)

[2] Baloch R 2015 *Ethnical Hacking and Penetration Testing Guide* (Boca Raton: CRC Press).

[3] Ross R 1996 *Stochastic processes* (New York: Wiley)

[4] Sarkka S 2017 *Bayesian Filtering and Smoothing* (Cambrige: Cambridge university press)

[5] Azarnova T V, Barkalov S A and Polukhin P V 2018 *Bulletin of South Ural state University. Series: Computer technology, control, radio electronics* **18(4)** 16-24

[6] Sirotkin A V 2006 SPIIRAS Proceedings **3(1)** 228-39

[7] Brooks S, Gelman A, Gones G and Meng Xi 2011 *Handbook of Markov Chain Mone Carlo* (Boca Raton:CRC Press)

[8] Winner N 1970 *Extrapolation,Interpolation and Smoothing of Stationary Series* (Massachusetts: MIT Press)

[9] Mikaeljan S V 2011 *Science and Edication* **10** 1-25

[10] Doucet A, Godsill S and Andrieu C 2000 *Statistics an Computing* **10(3)** 197-208

[11] Sobol I M 1973 *The numerical Monte Carlo methods* (Moscow: Nauka)