**Midterm Evaluation**

**CSE474**

**Review of Paper titled “DeviceMien: Network Device Behavior Modeling for Identifying Unknown IoT Devices”**

Reviewed by:

Imran Hossain Imon

ID: 19101277

Tasks:

1. **Which system is modeled and simulated in your selected paper?**

**Answer**: In the paper, the authors described a method for simulating device activity for both known and unknown devices using network traffic.

1. **Why was this simulation needed?**

**Answer:** The adoption of IoT devices is skyrocketing, making networks more attackable. To reduce the danger of an attack, network managers need improved tools for locating and verifying these devices. In order to provide useful feedback in device identification, especially when the device has not previously been observed, this research introduces a probabilistic framework. The simulation demonstrates that using the method, devices that have at least 50 samples observed may be identified with 100% accuracy after only 18.9 TCP-flow samples. By looking at the average number of flow classes seen throughout a set of data, it can also discriminate between two broad kinds of devices: IoT and Non-IoT. That is why this simulation is needed.

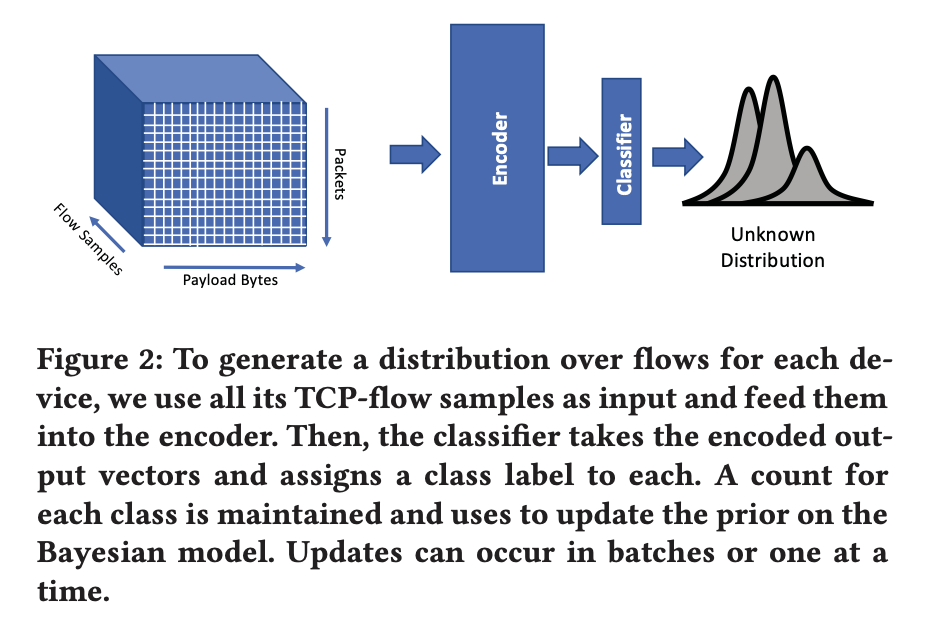
1. **Describe the system state variables of this system.**

**Answer:** There are 5 system state variables.

* **Entity:** IoT Devices cameras - (Nest Dropcam, Samsung SmartCam, Netatmo Welcome, Insteon Camera, TP-Link Day Night Cloud Cam- era, Withings Smart Baby Monitor), switches and triggers (iHome, TP-Link Smart Plug, Belkin Wemo Motion Sensor, Belkin Wemo Switch), hubs (Smart Things, Amazon Echo), air quality sensors (NEST Protect smoke alarm, Netatmo Weather station), electron- ics (Triby speaker, PIXSTAR Photoframe, HP Printer), healthcare devices (Withings Smart scale, Withings Aura smart sleep sensor, Blipcare blood pressure meter) and light bulbs (LiFX Smart Bulb). Non IoT Devices: laptop, mobile, tablets.
* **Attribute:** IoT, TCP, LSTM, RNN, Autoencoder, Bayesian Modeling, SVM.
* **Resource:** Probability, Clustering, Unsupervised Feature Learning.
* **Event, Activities and Delays:** Three processes are involved: (1) training; (2) model creation; and (3) execution and ranking. In the training phase (shown in Figure 1), develop a model that automatically picks up on how to describe packet flows, or collections of packets traveling between two sources. The clustering of all the representation samples discovered in the first step makes up the model-construction phase (Figure 1). Additionally, during this step, a classifier is trained to categorize the various flow types, and models are built for all recognized and tagged devices to represent that distribution. To learn a distribution of flows coming from new devices, the encoder and classifier are combined in the execution phase. Probabilistic measures of similarity are then used to compare and/or rank newly observed distributions with labeled distributions. (Figure 2 and Figure 3)

Diagram

Description automatically generated with low confidence

A picture containing arrow

Description automatically generatedFigure 1

1. **Classify the model of the system.**

**Answer:** There are three steps to our process. At first pre-process TCP flow data for each device in the first phase, known as the training phase, and then train a deep LSTM-Autoencoder network to extract a collection of representative features from the data itself. Then, adjust a clustering method that divides the features into the most distinct groups in accordance with the cluster silhouette score using a Bayesian hyper-parameter tuning framework. In other words, adjusting the clustering algorithm to increase the distance between the various clusters. The labels given to the clusters are then used to train a classifier.

The classifier is then used to create a distribution over these classes by categorizing each device's individual TCP-flow data in the following step of the technique. Then, using a Dirichlet prior on the parameters, model this distribution as a multinomial distribution. Finally, it produces a distribution, create a probabilistic model, and contrast it with the distributions of the other devices for each device which want to match.

1. **What questions did the author ask to analyze the problem?**

**Answer:** To assess the issue, the author posed two queries. First, determine whether or not the device's communication behavior is changing. This query was an attempt by the author to define the issue. Second, the author questioned whether the alterations fall within the parameters of previously noticed behavior patterns or are suggestive of malicious intent. The author made an effort to fully comprehend the issue.

1. **Describe the abstract model the authors came up with.**

**Answer:** The long-term, short-term (LSTM) neural network architecture and a stacked autoencoder architecture are the two types of networks that make up the feature-learning architecture. Sequences such as handwriting recognition, language modeling, and auditory modeling have all been effectively modeled using RNNs. RNNs have cyclical connections that send the output from the prior stage to the activation nodes. The RNN type known as the Long-Term Short-Term (LSTM) architecture is a variant of the common RNN. The memory blocks include unique multiplicative units called gates to regulate the information flow as well as memory cells with self-connections that store the network's temporal state. In the original architecture, each memory block had an input gate and an output gate. The flow of input activations into the memory cell is managed by the input gate. The output gate regulates how cell activations exit the cell and enter the remainder of the network.

An LSTM layer is connected to a stacked autoencoder network as input. The network is trained to reduce the disparity between the encoder's input and decoder's output. Back-propagation training is used to train the autoencoder model unsupervised. The network will produce more compact, higher-level semantic features by stacking multiple autoencoders, which has been demonstrated to capture useful latent feature representations and boost classifier performance in many applications.

The encoded vectors are clustered to obtain a distribution over the different TCP-flow types once the stacked, LSTM-Autoencoder network has been trained. After n observations have been made, to represent the distribution over k TCP-flow classes as a multinomial distribution. When rolling a k-sided die n times, the multinomial distribution simulates the likelihood of counts. As a result, need to describe the multinomial's parameters as a Dirichlet distribution and conjugate distribution of the multinomial, updating the Dirichlet parameters with each new observation.

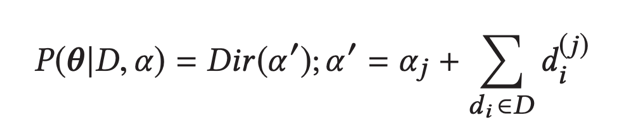
θ1,...,θk ∼Dir(α1,...,αk)

d1,...,dk ∼Mult(θ1,...,θk)

The posterior distribution had been determined by the application of Bayesian inference. According to the Bayes rule, the posterior is inversely proportional to the product of the likelihood of the data and our prior perception of the distribution over. Calculating the posterior is straightforward since it’s presumed that the distribution over types has a multinomial Dirichlet prior.

1. **Describe the mathematical model the authors came up with.**

**Answer:** The author only updates our prior with the count of each new observation of TCP-flow type for each new TCP connection that is observed for a specific device in order to compute the posterior distribution. The generative model also becomes straightforward to use as a result.



The distributions are comparing using the two-sample Kolmogorov-Smirnov test. Here Fn be defined as the empirical distribution function for n ordered observations Xi be defined as Fn (x)= 1/n\*∑ nI (X), where I is the indicator function n i=1 [−∞, x] i [−∞,x], equal to1 if Xi ≤ x and equal to 0 otherwise. The Kolmogorov- Smirnov statistic for F(x) I.

Dn = supx |Fn(x) − F(x)|

1. **How did the authors verify and validate their model? Write your suggestions for**

**improvement of the model.**

**Answer:** Two datasets were taken by the author. The first one is the University of New South Wales trail that is accessible to the general public. The second source is a MAC from a personal lab. With Gaussian Process priors, require employ Bayesian optimization. Then we can explore a model's parameters and scoring function in the dark using the Bayesian optimization with Gaussian Process (GP) priors’ technique. Finally, they demonstrate how the distributions of IoT and non-IoT devices differ in terms of behavior. This can aid in their decision-making in this situation and be used in conjunction with current strategies to offer a whole set of solutions that deal with the difficulties of defending networks from the onslaught of hazardous IoT and non-IoT devices equally.

1. **Name the tools and programs the authors used for the simulation.**

**Answer:** TCP Floe packet data, stacked autoencoder, Bayesian hyper-parameter

1. **What was the result of the simulation? What conclusions did the authors draw from their simulation output?**

**Answer:** The obvious trend of high scores along the diagonal is the first. This is to be expected because unlabeled models are created by randomly selecting a portion of the data distribution, thus they should all match their labeled counterparts. Examining the similarities between IoT devices both within and outside of classes, specifically the distributions of the (IoT, IoT) similarity score and the (Non-IoT, Non-IoT) similarity score. More than non-IoT devices, IoT devices resemble one another. The distribution within a class, however, is more skewed than the same-class distribution seen in the broader class comparison. A key drawback of supervised algorithms with weak (or no) priors is that they produce less precise conclusions with fewer observations. In our model and for all of our devices, the author utilizes a weak prior. To infer a better prior that could require fewer data to converge to a representative model, many strategies could be applied. Due to this, F1 scored 76% and Acc scored 61%.

At this point, the outcome is largely satisfactory in my perspective. The paper does, however, advise doing further simulation with more data in the future.

1. **Describe a new problem relevant to the paper you want to solve by modeling and**

**simulation.**

**Answer:** The purpose of the paper is to identify and identify unknown devices through network trafficking. But only identifying the unknown device is insufficient to reduce network assault. We must therefore categorize the devices that may be detrimental to the network. We must identify suspicious devices for both known and unidentified devices, categorize them as network vulnerabilities, and stop them from doing any harm. This can be accomplished by combining an RNN and a deep learning algorithm. In order to monitor the behavior of the devices, we may also define some parameters. If the values fall under our acceptable range, the device will be classified as vulnerable.