ON THE IDENTIFIABILITY OF MULTI-OBSERVER HIDDEN MARKOV MODELS

Hung X Nguyen and Matthew Roughan

School of Mathematical Sciences, The University of Adelaide, Australia
E-mail:hung.nguyen, matthew.roughan@adelaide.edu.au

ABSTRACT

Most large attacks on the Internet are distributed. As a result, such attacks are only partially observed by any one Internet service provider (ISP). Detection would be significantly easier with pooled observations, but privacy concerns often limit the information that providers are willing to share. Multi-party secure distributed computation provides a means for combining observations without compromising privacy.

In this paper, we show the benefits of this approach, the most notable of which is that combinations of observations solve identifiability problems in existing approaches for detecting network attacks.

Index Terms— Hidden Markov Models, Multiple Observers, Identifiability, Security, Networks

1. INTRODUCTION

The Internet allows communication between any two computers with network connections. This openness unfortunately has created many security problems. Even services such as SSH that are designed to be secure are vulnerable to brute force attacks. There are methods to defend against brute-force attacks, but we first need to detect the attack.

Hidden Markov Models (HMMs) have been used as an effective detector [1–3]. The actions of an attacker are modelled as a Markov process where each state of the process represents a step in the attack. Observations of network traffic are used to infer parameters of the hidden Markov process that generates the traffic. An anomaly/attack is detected if the model parameters match those of a typical SSH attack.

However, most large attacks are distributed. Hackers use botnets[1] to launch brute force SSH attacks from multiple sources. This allows hackers to hide their activities because each observer can only see a part of the attack. Pooled observations from multiple ISPs are essential to detect these distributed attacks.

Unfortunately, most current detection systems use data from only one Internet service provider (ISP) due to privacy concerns. ISPs that share data run the risk that their competitors will gain some advantage through the use of this data. Moreover, traffic data is protected by privacy legislation in many countries making it illegal to share this data with other parties. However, it has been recently shown that multi-party secure distributed computation can overcome this problem. In particular, recent work has shown how HMMs [4, 5] can be estimated, in particular for the Internet attack problem [6] from multiple observers, without private data being shared.

Previously, only the case where the observers all see the same distributions of observations was addressed. However, that case is unrealistic [6]. ISPs each have a different perspective. In this paper we show how the approach can be adapted to heterogeneous observers.

We also show in this paper that in a realistic HMM for these network attacks the model parameters are not identifiable from observations of a single ISP. Without identifiability, the estimated parameters of a HMM will be inaccurate no matter how much data is collected. However, observations from different ISPs with different perspectives can make the problem identifiable, and thus lead to accurate detection of attacks. This provides a strong incentive for ISPs to participate in such a collaborative detection algorithm.

2. HIDDEN MARKOV MODELS

A Markov chain is a sequence of random variables $\mathcal{Q} = q_1 \dots q_T$ with the Markov property: given the present state, the future and past states are independent. Consider a Markov chain with N possible states $\mathcal{S} = \{s_1, \dots, s_N\}$. If the states of the Markov process are not directly observed, but rather we see some outputs drawn from the set $\mathcal{V} = \{v_1, \dots, v_M\}$, which are probabilistically associated with the state of the Markov chain, the process is referred to as a Hidden Markov Model (HMM) [7]. A HMM is formally defined by the triplet:

- the initial probability $\pi = (\pi_1, \dots, \pi_N)$, where $\pi_i = \mathbb{P}(q_1 = s_i)$,
- $A = (a_{ij})_{N \times N}$, where $a_{ij} = \mathbb{P}(q_{t+1} = s_j | q_t = s_i)$, the time-independent state transition probability; and
- $B = (b_{ik})_{N \times M}$, where $b_{ik} = \mathbb{P}(O_t = v_k | q_t = s_i)$, the time-independent observation probability .

Most of the results discussed in this paper can be extended to the non-stationary case, but for brevity we restrict ourselves to the stationary case where the initial distribution π is also the stationary distribution, so the HMM is completely determined by $\boldsymbol{\lambda} = \{A, B\}$, and the parameters $\boldsymbol{\lambda}$ define a probability

[1] A botnet is a set of computers that are compromised and will follow instructions from the hacker.

measure $\mathbb{P}_\lambda$ on $\mathcal{V}^*$, the set of finite words from $\mathcal{V}$, including the empty word ϕ by

$$\mathbb{P}_\lambda(O_1 = v_{k_1}, \ldots, O_T = v_{k_T}) = \sum_{q_1,\ldots,q_T \in \mathcal{S}} \pi_{q_{i_1}} a_{i_1 i_2} b_{i_1 k_1} a_{i_2 i_3} b_{i_2 k_2} \ldots a_{i_{T-1} i_T} b_{i_T k_T}. \quad (1)$$

The standard definition of statistical *identifiability* is that

$$\mathbb{P}_\lambda(O_1, \ldots, O_T) = \mathbb{P}_{\tilde{\lambda}}(O_1, \ldots, O_T) \quad \Rightarrow \quad \boldsymbol{\lambda} = \tilde{\boldsymbol{\lambda}},$$

for all $T \in \mathbb{N}$, and all possible observations $O_1, \ldots, O_T$.

However, HMMs are not identifiable in the strict sense given above. The state labels in the Markov process are arbitrary, and so we can permute the states without changing the observation probabilities $\mathbb{P}_\lambda(O_1, \ldots, O_T)$. Also, we can always construct a HMM with additional states that is equivalent to $\boldsymbol{\lambda}$ [8]. Hence, a HMM is only ever identifiable with respect to a fixed number of states, and modulo permutations.

The joint probability $\mathbb{P}(O_{t+1} = v_k, q_{t+1} = s_j | q_t = s_i)$ plays an important role in the identifiability of a HMM. For an observation symbol $v_k \in \mathcal{V}$, let

$$M(k) = AB(k),$$

where $B(k) = \text{diag}\{b_{1k}, \ldots, b_{Nk}\}$. That is, $M(k)$ is an $N \times N$ matrix where each entry $m_{ij}(k) = \mathbb{P}(O_{t+1} = v_k, q_{t+1} = s_j | q_t = s_i)$. The observation probabilities $\mathbb{P}_\lambda(O_1, \ldots, O_T)$ can be expressed in terms of $M(\cdot)$ as

$$\mathbb{P}_\lambda(O_1, \ldots, O_T) = \boldsymbol{\pi} M(O_1) \ldots M(O_T) \boldsymbol{e}, \quad (2)$$

where e is the vector of length N with all 1 entries.

Two HMMs are said to be *equivalent* if they have the same observation probabilities. The following lemma from [9] provides sufficient conditions for two HMMs to be equivalent.

Lemma 1. *[9] Let $\boldsymbol{\lambda}$ and $\tilde{\boldsymbol{\lambda}}$ be the parameters of two HMMs with N and $\tilde{N}$ states respectively. If X and Y are $N \times \tilde{N}$ and $\tilde{N} \times N$ matrices respectively such that:*

$$\tilde{M}(k) = YM(k)X, \quad \forall v_k \in \mathcal{V}; \text{ and}$$

$$\tilde{\pi} = \pi X; \text{ and } \tilde{e} = Ye; \text{ and } XY = I_N;$$

then $\boldsymbol{\lambda}$ and $\tilde{\boldsymbol{\lambda}}$ are equivalent.

If two HMMs are equivalent, then neither is identifiable, but we need a more practical set of conditions. The identifiability of a HMM is closely related to the *rank* of $\mathbb{P}_\lambda(\cdot)$, which is defined below.

For the set of $2n$ words $w_1, \ldots, w_n, w'_1, \ldots, w'_n$ from $\mathcal{V}^*$, define a matrix $Q(w_1, \ldots, w_n, w'_1, \ldots, w'_n)$ whose $(i,j)th$ entry is $\mathbb{P}_\lambda(w_i, w'_j)$. Define *rank*$[\mathbb{P}_\lambda(\cdot)]$ to be the maximum of the rank of $Q(w_1, \ldots, w_n, w'_1, \ldots, w'_n)$, for all such words, for all n.

It is well-known [9] that all HMMs with N states have *rank*$[\mathbb{P}_\lambda(\cdot)] \leq N$. If the rank is N, then we say the HMM is *regular*. Finesso [8] showed that the regularity of a HMM can be determined in a finite number of operations. Petrie [10] proved the following sufficient conditions for identifiability (up to permutation of states) of discrete HMMs.

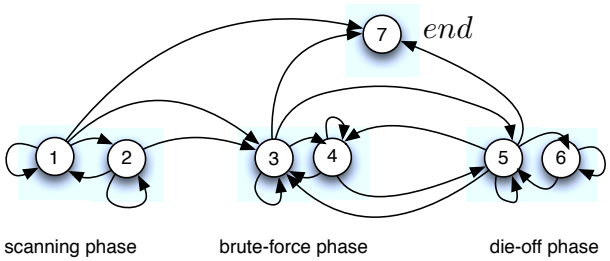

Fig. 1. The Markov chain for a typical SSH attack.

Lemma 2. *[10] The parameter $\boldsymbol{\lambda}$ of a HMM is identifiable if*

1. *the HMM is regular;*
2. *$M(k)$ is invertible , $\forall v_k \in \mathcal{V}$; and*
3. *$\exists v_k \in \mathcal{V}$ such that $b_{ik}, (i = 1, 2, \ldots, N)$, are distinct.*

3. SINGLE OBSERVER HMM FOR SSH ATTACKS

Typical SSH brute-force attacks often go through three phases [3]. In the first phase – *scanning* – the attacker scans the target network for vulnerable SSH services. In the second phase – *brute-force* – the attacker initiates a brute-force user/password dictionary based attack on the vulnerable hosts. In the last *die-off* phase, compromised hosts communicate with the attackers and wait for new instructions. During each phase, the attackers alternate between an active and inactive state to make detection more difficult. For the attacks observed in [3], when active, the average number of packets per flow is 1.5 in the scanning phase, 11 in the attack phase, and 1.5 in the die-off phase. Note that these values are only examples. A Markov model shown in Figure 1, with seven states, is used to represent the various stages of an attack. The observations are the numbers of packets per flow with observation probabilities

$$B = \begin{pmatrix} \epsilon & 1-2\epsilon & \epsilon & 1-2\epsilon & \epsilon & 1-2\epsilon & 1-2\epsilon \\ 1-2\epsilon & \epsilon & \epsilon & \epsilon & 1-2\epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & 1-2\epsilon & \epsilon & \epsilon & \epsilon & \epsilon \end{pmatrix}^T,$$

where ϵ is the probability of measurement error.

Unfortunately, this model is not uniquely identifiable as the parameters of the HMM do not satisfy condition 3 in Lemma 2. Indeed, as states 6 and 7 have the same observation probabilities, we can use the construction in [11] to obtain a set of equivalent HMMs as follows. Take

$$Y = P \begin{bmatrix} I & 0 \\ 0 & F \end{bmatrix}, \; X = Y^{-1},$$

where P is a permutation matrix, $F \in \mathbb{R}^{2\times 2}$ is a nonsingular matrix with $F\boldsymbol{e} = \boldsymbol{e}$, and I is a 5×5 identity matrix. Then define new matrices

$$\tilde{B} = B \text{ and } \tilde{A}\tilde{B}(k) = YAB(k)X, \; \forall v_k \in \mathcal{V},$$

and we can verify that the two HMMs are equivalent by showing that the conditions in Lemma 1 are satisfied. Where more than one model fits the observations, the estimated parameters can be badly in error, no matter how much data we collect.

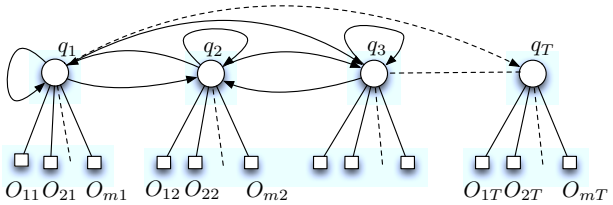

Fig. 2. A HMM with multiple observers.

4. MULTI-OBSERVER HMM

We now consider the case where multiple ISPs monitor the same distributed attack and combine their data using the privacy-preserving techniques described in [6].

We model the attack as before, by a Markov process with transition matrix A. However, there are now m ISPs that make observations of the same underlying Markov process. Each ISP makes its own observations of the attack and these observations are secret. Each ISP could treat the problem as a private single-observer HMM with parameter $\boldsymbol{\lambda}^{(j)} = \{A, B^{(j)}\}$, but when the ISPs combine their measurements to jointly infer the under-lying common hidden Markov model, we have a *multi-observer HMM*.

We shall assume that the observations of the different ISPs are independent. That is a natural assumption, as dependence would weaken the need for multiple observers, or privacy. Assume also that the set of possible observations $\mathcal{V}$ is the same for all ISPs. However, each ISP j has its own observation probability given by the matrix $B^{(j)} = \{b_{ik}^{(j)}\}$. These differ because each ISP has a different perspective on the attack.

In this new model, the observation set is $\mathcal{V}^m$. Denote the observations

$$\mathcal{O} = \{\mathcal{O}_1, \ldots, \mathcal{O}_T\},$$

where each element $\mathcal{O}_t$ is a vector $\mathcal{O}_t = \{O_{1t}, \ldots, O_{mt}\}$ of observations at time t from each ISP. The sequence of T observations that ISP j makes is denoted as $\mathcal{O}^{(j)} = \{O_{j1}, \ldots, O_{jT}\}$. An example is given in Figure 2.

The probability of a set of observations at time t, conditional on the state of the Markov process, is given by

$$\mathbb{P}(\mathcal{O}_t | q_t = s_i) = \prod_{j=1}^{m} \mathbb{P}(O_{jt} = v_{k_j} | q_t = s_i) = \prod_{j=1}^{m} b_{ik_j}^{(j)}. \quad (3)$$

The multi-observer HMM has transition matrix A and the observation probabilities given in (3). Its parameter set is therefore $\boldsymbol{\lambda}_{multi} = \{A, \{B^{(j)}\}\}$. From T observations of the multi-observer HMM, $\{\mathcal{O}_1, \ldots, \mathcal{O}_T\}$, we can infer the matrix A and the observation probabilities $\mathbb{P}(\mathcal{O}_t | q_t = s_i)$ using the Baum-Welch algorithm [7]. We presented a privacy-preserving protocol for the Baum-Welch algorithm in [6], in the homogenous case where all ISPs have the same observation probabilities. We have extended this protocol to the heterogeneous case (details omitted for brevity), and we evaluate the accuracy of this extension in the next section.

In addition, in the homogenous case [6], the identifiability of the multi-observer HMM was the same as for the single observer case. However, when the observers are heterogeneous, the multi-observer problem may be identifiable, even if the single observer problem is not, as shown below.

Define for every word $w = v_1 \cdots v_k \in \mathcal{V}^*$

$$M(w) \doteq M(v_1) \ldots M(v_k);\ g(w) \doteq \pi M(w);\ h(w) \doteq M(w)e.$$

For the set of $2N$ words $w_1, \ldots, w_N, w'_1, \ldots, w'_N$, let G and H be two matrices of size $N \times N$ where the i-th row of G is $g(w_i)$ and the j-th column of H is $h(w'_j)$. Similar notations are used for the multi-observer HMM where each observation is a vector of m individual observations. From [8],

$$Q(w_1, \ldots, w_N, w'_1, \ldots, w'_N) = G(w_1, \ldots, w_N) H(w'_1, \ldots, w'_N).$$

Lemma 3. *The multi-observer HMM is regular if*

- $\exists j \leq m$ *such that HMM* $\boldsymbol{\lambda}^{(j)}$ *is regular; and*
- *the observation probabilities* $b_{ik}^{(j)} > 0$ *for all* $i \leq N, k \leq M, j \leq m$.

Proof. Without loss of generality, assume that the single observer HMM $\boldsymbol{\lambda}^{(1)}$ is regular. Thus there exist 2N words $w_1, \ldots, w_N,\ w'_1, \ldots, w'_N$, each of length $n_i = |w_i|$, such that $Q(w_1, \ldots, w_N, w'_1, \ldots, w'_N)$ is non-singular [8]. Denote the symbols of the word w_i by $w_i = v_{k_{i,1}} \ldots v_{k_{i,n_i}}$. As $Q(w_1, \ldots, w_N, w'_1, \ldots, w'_N)$ is non-singular, the matrix

$$G(w_1, \ldots, w_N) = \left[\pi A^{n_i} \prod_{l=1}^{n_i} B^{(1)}(k_{i,l}) \right]_{1 \leq i \leq N} \quad (4)$$

is also non-singular, i.e., its rows are linearly independent.

In the multi-observer HMM, consider the $2N$ words $y_1, \ldots, y_N, y'_1, \ldots, y'_N$ where each word y_i has n_i symbols in which the l-th symbol is a vector of m components with the first component being $v_{k_{i,l}}$ and the others v_1, i.e., $y_i = \{v_{k_{i,1}}, v_1, \ldots, v_1\} \ldots \{v_{k_{i,n_i}}, v_1, \ldots, v_1\}$. The i-th row of the matrix $G(y_1, \ldots, y_N)$ is

$$\begin{aligned} g(y_i) = \pi M(y_i) &= \pi \textstyle\prod_{l=1}^{n_i} \left(A B^{(1)}(k_{i,l}) \prod_{j=2}^{m} B^{(j)}(1) \right) \\ &= \pi A^{n_i} \left(\textstyle\prod_{l=1}^{n_i} B^{(1)}(k_{i,l}) \right) \left(\prod_{j=2}^{m} B^{(j)}(1) \right)^{n_i}. \quad (5) \end{aligned}$$

Since the rows of the matrix in (4) are linearly independent and $b_{ik}^{(j)} > 0$ for all i, j, k, the rows $g(y_i)$ of $G(y_1, \ldots, y_N)$ given in (5) are therefore also linearly independent. Thus, $G(y_1, \ldots, y_N)$ is non-singular.

Similarly, we can prove that $H(y'_1, \ldots, y'_N)$ is non-singular. Hence, $Q(y_1, \ldots, y_N, y'_1, \ldots, y'_N)$ is non-singular and the multi-observer HMM is regular. □

Theorem 1. *The matrix A in the multi-observer HMM is identifiable if*

1. *at least one of the HMMs* $\{\boldsymbol{\lambda}^{(1)}, \ldots, \boldsymbol{\lambda}^{(m)}\}$ *is regular;*
2. *A is a non-singular stochastic matrix;*

3. $b_{ik}^{(j)} > 0$ *for all* $i \leq N, k \leq M, j \leq m$*;*

4. $\exists v_{k_1} \dots v_{k_m} \in \mathcal{V}$ *such that* $\prod_{j=1}^{m} b_{ik_j}^{(j)}$ *are distinct for* $i = 1, 2, \dots, N$

Proof. We prove that the multi-observer HMM satisfies the Petrie's conditions in Lemma 2. From Condition 1 and 3 and Lemma 3, the multi-observer HMM is regular.

The second of Petrie's conditions that $M(\{v_{k_1}, \dots, v_{k_m}\})$ is invertible for all symbols $\{v_{k_1}, \dots, v_{k_m}\} \in \mathcal{V}^m$ follows from the fact that $M(v_{k_1}, \dots, v_{k_m}) = A \prod_{j=1}^{m} B^{(j)}(k_j)$ and that the matrices A and $B^{(j)}(k_j)$ are non-singular.

The last of Petrie's conditions comes from Condition 4 that $\exists v_{k_1} \dots v_{k_m} \in \mathcal{V}$ such that $\prod_{j=1}^{m} b_{ik_j}^{(j)}$ are distinct. □

Note that almost all HMMs are regular and satisfy the first two conditions [10]. If the single observer HMMs are regular but not identifiable, the multi-observer HMM is identifiable under mild conditions (condition 3 and 4 in Theorem 1) on the individual observation probability.

5. EVALUATION

We have extended the implementation of privacy preserving protocols in [6] to the multi-observer HMM with heterogenous observation probabilities discussed in this paper. The code (written in Python) is available at *www.hxnguyen.net*. We evaluate the accuracy of this implementation, especially the effect of identifiability by running simulations for the SSH attack HMM in Section 4 with m observers. The transition matrix A is as in [3]. For each observer, we create a random observation matrix of size 7×3 with the same observation probabilities for states 6 and 7 as in Section 3.

Using 128 bit encryption keys for privacy preserving computation, we compare the errors as we increase the number of participants m in the protocol. We apply the secure Baum-Welch algorithm to infer the transition matrix A with $T = 100$ samples. We compare errors in the estimates by calculating the Mean Squared Error (MSE) over the matrix A.

The resultant MSE is shown in Figure 3. The plot shows that there is a substantial increase in accuracy when two parties collaborate, but that the marginal improvement lessens with increasing numbers of participants in the protocol. This dramatic improvement can be explained by the fact that the matrix A is unidentifiable with one observer but is identifiable with multiple observers.

6. CONCLUSION AND FUTURE WORK

In this paper we have shown that collaboration between multiple parties can improve the quality of estimates provided by HMMs. More importantly, using privacy preserving techniques the parties can collaborate without revealing private data to each other. In the context of ISPs, this would mean that multiple ISPs can help each other detect network problems without running the risk of exposing critical data. Our future work includes finding the necessary conditions for identifiability of multi-observer HMMs.

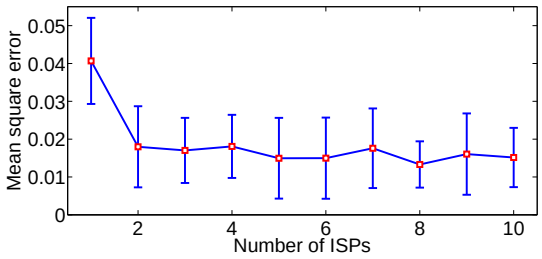

Fig. 3. MSE of the transition probabilities A.

7. REFERENCES

[1] Charles V. Wright, Fabian Monrose, and Gerald M. Masson, "On inferring application protocol behaviors in encrypted network traffic," *J. Mach. Learn. Res.*, vol. 7, pp. 2745–2769, December 2006.

[2] Davide Ariu, Giorgio Giacinto, and Roberto Perdisci, "Sensing attacks in computers networks with Hidden Markov Models," in *Proceedings of the Machine Learning and Data Mining in Pattern Recognition - MLDM*, 2007, pp. 449–463.

[3] Claudio Bartolini, Luciano Gaspary, Anna Sperotto, Ramin Sadre, Pieter-Tjerk de Boer, and Aiko Pras, "Hidden Markov Model modeling of SSH brute-force attacks," in *Integrated Management of Systems, Services, Processes and People in IT*. 2009, pp. 164–176, Springer Berlin / Heidelberg.

[4] Manas Pathak, Shantanu Rane, Wei Sun, and Bhiksha Raj, "Privacy preserving probabilistic inference with Hidden Markov Models," in *Proc. of IEEE ICASSP 2011*, 2011.

[5] P. Smaragdis and M. Shashanka, "A framework for secure speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1404 –1413, May 2007.

[6] Hung X Nguyen and Matthew Roughan, "Multi-observer privacy-preserving Hidden Markov Models," Tech. Rep., University of Adelaide, 2011, `http://www.hxnguyen.net/papers/TR_HMM.pdf`.

[7] Lawrence R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[8] Lorenzo Finesso, "Consistent estimation of the order for Markov and hidden Markov Chains," *Ph.D. thesis, Maryland University*, 1991.

[9] Edgar J. Gilbert, "On the identifiability problem for functions of finite Markov chains," *The Annals of Mathematical Statistics*, vol. 30, no. 3, pp. 688–697, 1959.

[10] T. Petrie, "Probabilistic functions of finite state Markov chains," *The Annals of Mathematical Statistics*, vol. 40, no. 1, pp. 97–115, 1969.

[11] Bart Vanluyten, Jan C. Willems, and Bart De Moor, "Equivalence of state representations for hidden Markov models," *Systems & Control Letters*, vol. 57, pp. 410–419, 2008.