A Tamper-Free Semi-Universal Communication System for Deletion Channels

Shahab Asoodeh[1], Yi Huang[2], and Ishanu Chattopadhyay[3]

Abstract

We investigate the problem of reliable communication between two legitimate parties over deletion channels under an active eavesdropping (aka jamming) adversarial model. To this goal, we develop a theoretical framework based on probabilistic finite-state automata to define novel encoding and decoding schemes that ensure small error probability in both message decoding as well as tamper detecting. We then experimentally verify the reliability and tamper-detection property of our scheme.

I. Introduction

The deletion channel is the simplest point-to-point communication channel that models synchronization errors. In the simplest form, the inputs are either deleted independently with probability δ or transmitted over the channel noiselessly. As a result, the length of channel output is a random variable depending on δ. Surprisingly, the capacity of deletion channel has been one of the outstanding open problems in information theory [1]. A random coding argument for proving a Shannon-like capacity result for deletion channel (in general for all channels with synchronization errors) was given by Dobrushin [2] which is recently improved by Kirsch and Drinea [3] to derive several lower bounds. Readers interested in most recent results on deletion channels are referred to the recent survey by Mitzenmacher [4] that provides a useful history and known results on deletion channels.

As the problem of computing capacity of deletion channels is infamously hard, we focus on another problem in deletion channels. In this paper, we study the behavior of the deletion channel under an *active* eavesdropper attack. Secrecy models in information theory literature, initiated by Yamamoto [5], assume that there exists a *passive* eavesdropper who can observe the symbols being transmitted over the channel. The objective is to design a pair of (randomized) encoder and decoder such that the message is decoded with asymptotically vanishing error probability at the legitimate receiver while ensuring that the eavesdropper gains negligible information about the message. In all secrecy models (see, e.g., [6]–[12]) the crucial assumption is that the eavesdropper can neither jam the communication channel between legitimate parties nor can she modify any messages exchanged between them. However, in many practical scenarios, the eavesdropper can potentially change the channel, for instance, add stronger noise to change the crossover probability of a binary symmetric channel or the deletion probability of a deletion channel.

In our adversarial model, we assume that two parties (say Alice and Bob) wish to communicate over a public deletion channel while an eavesdropper (say Eve) can potentially tamper the statistics of the channel. We focus on deletion channel and assume that Eve can have possibly more bits deleted, and hence increases the deletion probability of the channel. The objective is to allow a reliable communication between Alice and Bob (with vanishing error probability) regardless of the eavesdropper's action. To this goal, we design (i) a randomized encoder using probabilistic finite-state automata which, given a fixed message, generates a random vector as the channel input and (ii) a decoder which generates an estimate of the message *only* when the channel is not tampered. In case the channel is indeed tampered, the decoder can declare it with asymptotically small Type I and Type II error probabilities. It is worth mentioning that the *rate* of our scheme is (almost) zero and hence we do not intend to study capacity of deletion channels.

Unlike the classical channel coding where the set of all possible channel inputs (aka, codebook) must be available at the decoder, our scheme requires that only the set of PFSA's used in the encoder to be available at the decoder. This model, that we call *semi-universal*, is contrasted with *universal* channel coding [13] where neither channel statistics nor codebook are known and the decoder is required to find the *pattern* of the message.

The rest of the paper is organized as follows. In Section II, we discuss briefly the notion of PFSA and its properties required for our scheme. Section III specifies the channel model, encoder, decoder, and different error events. In Section IV, we discuss the effects of deletion channels on PFSA. Section V concerns the thoeretical aspects of our coding scheme and Section VI contains several experimental results.

Notation We use calligraphic uppercase letters for sets (e.g. $\mathcal{S}$), sans serif font for functions (e.g. T), uppercase letters for matrices (e.g. Γ), bold lower case letters for vectors (e.g. $\mathbf{v}$). Throughout, we use g to denote a PFSA and s and x to denote its state and symbol, respectively. We use $\mathbf{x}^n = x_1 \dots x_n$ for a sequence of symbols or interchangeably, $\mathbf{x}$ if its size is clear in context. Also, $\mathbf{v}_i$ for ith entry of vector $\mathbf{v}$, $A_{i,\cdot}$ and $A_{\cdot,j}$ for the ith row or column of the matrix A, respectively. We use $(a_x)_{x\in\mathcal{X}}$ to denote a vector with the entry indexed by x and $(\mathbf{a}_x)_{x\in\mathcal{X}}$ a matrix with the column indexed by x. Finally, $\mathbf{x}^i = x_1 x_2 \dots x_i$.

[1]Computation Institute and Institute of Genomics and System Biology, The University of Chicago, Chicago, IL 60637 shahab@uchicago.edu
[2] The University of Chicago, Chicago, IL yhuang10@uchicago.edu
[3]Computation Institute, Chicago, IL ishanu@uchicago.edu

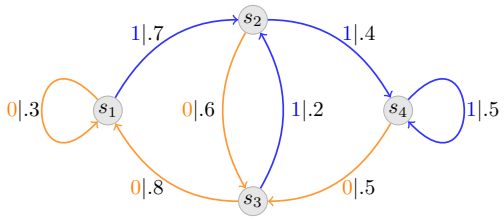


Fig. 1. A PFSA with $\mathcal{S} = \{s_1, s_2, s_3, s_4\}$ and $\mathcal{X} = \{0, 1\}$.

II. Probabilistic finite state automata

In this section, we introduce a new measure of similarity between two vectors. To do this, we first need to define probabilistic finite-state automata (PFSA).

Definition 1 (PFSA)**.** *A probabilistic finite-state automaton is a quadruple* $(\mathcal{S}, \mathcal{X}, \mathsf{T}, \mathsf{P})$*, where* $\mathcal{S}$ *is a finite state space,* $\mathcal{X}$ *is a finite alphabet with* $K = |\mathcal{X}|$*,* $\mathsf{T} : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ *is the state transition function, and* $\mathsf{P} : \mathcal{S} \times \mathcal{X} \to [0, 1]$ *specifies the conditional distribution of generating a symbol conditioned on the state.*

In fact, a PFSA is a directed graph with a finite number of vertices (i.e., states) and directed edges emanating from each vertex to the other. An edge from state $s_1 \in \mathcal{S}$ to state $s_2 \in \mathcal{S}$ is specified by two labels: (i) a symbol $x \in \mathcal{X}$ that updates the current state from s_1 to s_2, that is, $\mathsf{T}(s_1, x) = s_2$, and (ii) the probability of generating x when the system resides in state s_1, i.e., $\mathsf{P}(s_1, x)$. For instance, $\mathsf{P}(s_1, 1) = 0.7$ in the PFSA described in Fig. 1, thus, the system residing in states s_1 evolve to state s_2 with probability 0.7 and it generates symbol 1. Clearly, $\sum_{x\in\mathcal{X}} \mathsf{P}(s, x) = 1$ for all $s \in \mathcal{S}$. Given two symbols x_1 and x_2, one can define the transition function for the concatenation x_1x_2 as $\mathsf{T}(s, x_1x_2) = \mathsf{T}(\mathsf{T}(s, x_1), x_2)$. Letting $\mathcal{X}^*$ denote the set of all possible concatenation of finitely many symbols from $\mathcal{X}$, one can easily proceed to define $\mathsf{T}(s, \mathbf{x})$ as above for each $\mathbf{x} \in \mathcal{X}^*$ and $s \in \mathcal{S}$. We say that a PFSA is *strongly connected* if for any pair of distinct states s_i and s_j, there exists a sequence $\mathbf{x} \in \mathcal{X}^*$ such that $\mathsf{T}(s_i, \mathbf{x}) = s_j$. Let $\mathcal{G}$ be the set of all strongly connected PFSAs. The significance of strongly connected PFSAs is that their corresponding Markov chains (i.e., the Markov chain with state space $\mathcal{S}$ and transition matrix $P = [P(i, j)]_{|\mathcal{S}|\times|\mathcal{S}|}$ whose entry is $P(i, j) = \sum_{x\in\mathcal{X}:\mathsf{T}(s_i,x)=s_j} \mathsf{P}(s_i, x)$) has a unique stationary distribution (thus initial state can be assumed to be irrelevant).

Definition 2 (Γ-expression for PFSA)**.** *We notice that a PFSA* g *is uniquely determined by* $\mathbf{\Gamma}_g = (\Gamma_{g,x})_{x\in\mathcal{X}}$ *given by*

$$(\Gamma_{g,x})_{i,j} = \begin{cases} \mathsf{P}_g(s_i, x), & \mathsf{T}_g(s_i, x) = s_j, \\ 0, & \textit{otherwise}. \end{cases}$$

The state-to-state transition matrix P_g *is defined as*

$$P_g = \sum_{x\in\mathcal{X}} \Gamma_{g,x}, \tag{1}$$

and the state-to-symbol transition matrix $\widetilde{P}_g$ *is given by*

$$\widetilde{P}_g = (\Gamma_{g,x}\mathbf{1}_{|\mathcal{S}|})_{x\in\mathcal{X}},$$

where $\mathbf{1}_n$ *is the length-*n *all-one vector.*

For the PFSA illustrated in Fig. 1, we have

$$\Gamma_{g,0} = \begin{pmatrix} .3 & 0 & 0 & 0 \\ 0 & 0 & .6 & 0 \\ .8 & 0 & 0 & 0 \\ 0 & 0 & .5 & 0 \end{pmatrix}, \quad \Gamma_{g,1} = \begin{pmatrix} 0 & .7 & 0 & 0 \\ 0 & 0 & 0 & .4 \\ 0 & .2 & 0 & 0 \\ 0 & 0 & 0 & .5 \end{pmatrix},$$

$$P_g = \begin{pmatrix} .3 & .7 & 0 & 0 \\ 0 & 0 & .6 & .4 \\ .8 & .2 & 0 & 0 \\ 0 & 0 & .5 & .5 \end{pmatrix}, \quad \text{and} \quad \widetilde{P}_g = \begin{pmatrix} .3 & .7 \\ .6 & .4 \\ .8 & .2 \\ .5 & .5 \end{pmatrix}.$$

Definition 3 (Generalized PFSA)**.** *Generalized PFSA is a PFSA* g *whose* $\Gamma_{g,x}$ *can have more than one non-zero (positive) entries. In this case, we still have*

$$\left(\Gamma_{g,x}\mathbf{1}_{|\mathcal{S}|}\right)_i = \mathsf{P}_g(s_i, x).$$

However, $\mathsf{T}(s_i, x)$ *might not be deterministic, and instead it is a probability distribution.*

Shannon [14] appears to be the first one who made use of PFSAs to describe stationary and ergodic sources. Given $g \in \mathcal{G}$, first a state s_1 is chosen randomly according to the stationary distribution, then a symbol x_1 is generated with probability $\mathsf{P}(s_1, x_1)$ which takes the system from state s_1 to state s_2. A new symbol x_2 is then generated with probability $\mathsf{P}(s_2, x_2)$. Letting this process run for n time units, we obtain a sequence $x_1, x_2, \ldots, x_n$. In this case, we say that $x_1, x_2, \ldots, x_n$ is a *realization* of g. According to Shannon, each state s_i captures the "residue of influence" of the preceding symbol x_{i-1} on the system.

For $\mathbf{x} \in \mathcal{X}^*$, we denote by $\mathbf{x} \leftarrow g$ the fact that $g \in \mathcal{G}$ generates $\mathbf{x}$.

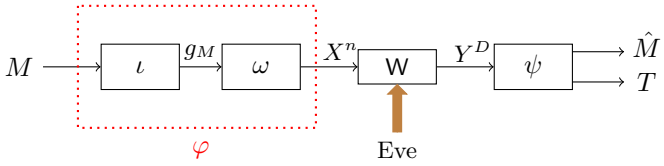


Fig. 2. A communication system with an active eavesdropper

III. System Model and Setup

Suppose Alice has a message M which takes value in a finite set $\mathcal{M} \coloneqq \{1, 2, \ldots, |\mathcal{M}|\}$ and seeks to transmit it reliably to Bob over a deletion channel $\mathsf{W}(\delta)$ with deletion probability $\delta \in [0, 1]$. The communication channel is assumed to be public, that is, an active eavesdropper, say Eve, can access and possibly tamper the channel. For simplicity, we assume that Eve may delete extra bits and thus changing the channel from $\mathsf{W}(\delta)$ to $\mathsf{W}(\delta')$ with $\delta' \geq \delta$.

The objective is to design a pair of encoder φ and decoder ψ that enables Alice and Bob to reliably communicate over $\mathsf{W}(\delta)$ *only* when he is ensured that the channel is not tampered. In classical information theory, the decoder must be tuned with the channel statistics. Hence, reliable communication occurs only when Bob knows the deletion probability δ. However, Eve might have tampered the channel and increased deletion probability to δ', and since Bob's decoding policy was tuned to δ, this might cause a decoding error –regardless of Bob's decoding algorithm. Therefore, reliability of the decoding must be always conditioned on the fact that the channel has not been tampered during communication.

Motivated by this observation, we propose the following coding scheme. We first propose a two-step encoder: each message $M = m$ is first sent to a function $\iota : \mathcal{M} \to \mathcal{G}_{\mathcal{M}}$ which maps m to a PFSA g_m in $\mathcal{G}_{\mathcal{M}} \coloneqq \{g_1, \ldots, g_{|\mathcal{M}|}\}$, then another function $\omega : \mathcal{G}_{\mathcal{M}} \to \mathcal{X}^n$ generates $\mathbf{y}^n$ a realization of PFSA g_m and sends it over the memoryless channel $\mathsf{W}(\delta)$. Therefore, the encoder function $\varphi : \mathcal{M} \to \mathcal{X}^n$ is the composition $\iota \circ \omega$ (see Fig. 2). Unlike the classical setting, Bob need not know the set of all channel inputs $\mathbf{y}^n$ for each $m \in \mathcal{M}$ (aka codebook). Instead, we assume Bob knows $\mathcal{G}_{\mathcal{M}}$ (thus the name *semi-universal scheme*). The output of the channel $\mathbf{x}^D$ is an $\mathcal{X}$-valued random vector whose length D is a binomial random variable $\mathsf{Bin}(n, 1-\delta)$ (corresponding to how many elements of $\mathbf{y}^n$ are deleted). Upon receiving $\mathbf{x}^D$, Bob applies $\psi : \mathcal{X}^* \to \mathcal{M} \times \{0, 1\}$ to generate $\psi(\mathbf{x}^D) = (\hat{M}, T)$ where $\hat{M}$ is an estimate of Alice's message and T specifies whether or not the channel has been tampered. He then declares $\hat{M}$ as the message only when $T = 0$. Therefore, the goal is to design (φ, ψ) such that for sufficiently large n

$$\Pr(T = 0 \mid \text{channel is tampered}) + \Pr(T = 1 \mid \text{channel is not tampered}) < \varepsilon, \tag{2}$$

and simultaneously

$$\Pr(M \neq \hat{M} | T = 0) \leq \varepsilon, \tag{3}$$

for any uniformly chosen message $M \in \mathcal{M}$. We say that the reliable tamper-free communication is possible if (2) and (3) hold simultaneously for any $\varepsilon > 0$.

IV. PFSA through deletion channel

In this section, we study the channel effect on PFSA's by monitoring the change of the likelihood of $\mathbf{x}$ being generated by a PFSA at the channel output. To do this, we first study the likelihood when $\delta = 0$ in Section IV-A, and then move on to the case of positive δ in Section IV-B. One of the main results in this section is to show that the output of $\mathsf{W}(\delta)$ (i.e., $\mathbf{x}^D$) can be equivalently generated by a *generalized* PFSA $g(\delta)$ whose Γ and state-to-state transition matrix follow simple closed forms (cf. Theorem 1). In section IV-C, we discuss some basic properties of $g(\delta)$ that will be useful for later development. We conclude this section by introducing the class M2 of PFSAs which is closed under deletion. For notational brevity, we remove the subscript g when it is is clearly understood from context.

A. PFSA over $\mathsf{W}(0)$: no deletion

Let a sequence of symbols $\mathbf{x} = x_1 \ldots x_n \in \mathcal{X}^*$ be given. We define $p_g(\mathbf{x})$ (or simply $p(\mathbf{x})$) to be the probability that g generates $\mathbf{x}$. Then we have

$$p(\mathbf{x}^n) = p(x_1) p\left(x_2 | \mathbf{x}^1\right) \cdots p\left(x_n | \mathbf{x}^{n-1}\right),$$

where $p\left(x_i|\mathbf{x}^{i-1}\right)$ is the conditional probability of g generating x_i given that g generated $\mathbf{x}^{i-1}$. It is clear from section II that

$$\begin{aligned}
\mathbf{p}_0 &= \mathbf{p} \qquad (4)\\
p(x_1) = \left(\mathbf{p}_0^T \widetilde{P}\right)_{x_1},\ \mathbf{p}_1^T &= \frac{\mathbf{p}_0^T \Gamma_{x_1}}{\left\|\mathbf{p}_0^T \Gamma_{x_1}\right\|_1},\\
p\left(x_2|x_1\right) = \left(\mathbf{p}_1^T \widetilde{P}\right)_{x_2},\ \mathbf{p}_2^T &= \frac{\mathbf{p}_1^T \Gamma_{x_2}}{\left\|\mathbf{p}_1^T \Gamma_{x_2}\right\|_1},\\
&\vdots\\
p\left(x_{n-1}|\mathbf{x}^{n-2}\right) = \left(\mathbf{p}_{n-2}^T \widetilde{P}\right)_{x_{n-1}},\ \mathbf{p}_{n-1}^T &= \frac{\mathbf{p}_{n-2}^T \Gamma_{x_{n-1}}}{\left\|\mathbf{p}_{n-2}^T \Gamma_{x_{n-1}}\right\|_1},
\end{aligned}$$

and finally, $p\left(x_n|\mathbf{x}^{n-1}\right) = \left(\mathbf{p}_{n-1}^T \widetilde{P}\right)_{x_n}$, where T denotes matrix transpose.

It is clear from the above update rule that any sequence x^{n} induces two probability distribution: one on the state space $\mathcal{S}$, i.e., $\mathbf{p}_n$ and the other one on $\mathcal{X}$. Let denote the former by $\mathbf{p}_g(\mathbf{x})$ and the latter by $\mathcal{D}_g(\mathbf{x})$. Update rules in (4) imply that $\mathcal{D}_g(\mathbf{x}) = \mathbf{p}_g^T(\mathbf{x})\widetilde{P}_g$ and $\mathbf{p}_g^T(\mathbf{x}x) \propto \mathbf{p}_g^T(\mathbf{x})\Gamma_{g,x}$. More precisely, since

$$\left\|\mathbf{p}_g^T(\mathbf{x})\Gamma_{g,x}\right\|_1 = \mathbf{p}_g^T(\mathbf{x})\Gamma_{g,x}\mathbf{1}_{|\mathcal{S}|} = \mathbf{p}_g^T(\mathbf{x})\left(\widetilde{P}_g\right)_{\cdot,x} = \left(\mathbf{p}_g^T(\mathbf{x})\widetilde{P}_g\right)_x = p(x|\mathbf{x}),$$

we have

$$p^T(x|\mathbf{x})\mathbf{p}_g(\mathbf{x}x) = \mathbf{p}_g^T(\mathbf{x})\Gamma_{g,x}. \qquad (5)$$

We also call $\mathcal{D}_g(\mathbf{x}) = \left(p_g(x|\mathbf{x})\right)_{x\in\mathcal{X}}$ the *symbolic derivative* of g induced by $\mathbf{x}$.

B. PFSA over $W(\delta)$: deletion with probability $\delta > 0$

Now we move forward to investigate the effect of deletion probability on PFSA transmission. The following result is a ket for our analysis.

Theorem 1. *Let $\mathbf{y} \leftarrow g$ be a channel input and $\mathbf{x}$ be a channel output with positive deletion probability δ. Then $\mathbf{x} \leftarrow g(\delta)$, where $g(\delta)$ is a generalized PFSA identified by $\Gamma_{g,x,\delta} = Q(P,\delta)\Gamma_{g,x}$ for all $x \in \mathcal{X}$, where P is the state-to-state transition matrix of g and Q is as defined in (6).*

Proof. Assume Bob has observed $\mathbf{x}^{i-1}$. Then we have

$$\begin{aligned}
p(x_i|\mathbf{x}^{i-1}) &= (1-\delta)\left(\mathbf{p}_{i-1}^T\widetilde{P}\right)_{x_i} + \delta(1-\delta)\left(\mathbf{p}_{i-1}^T P\widetilde{P}\right)_{x_i} + \delta^2(1-\delta)\left(\mathbf{p}_{i-1}^T P^2\widetilde{P}\right)_{x_i} + \cdots\\
&= (1-\delta)\left(\mathbf{p}_{i-1}^T\left(\sum_{i=0}^{\infty}\delta^i P^i\right)\widetilde{P}\right)_{x_i}\\
&= \left(\mathbf{p}_{i-1}^T Q(P,\delta)\widetilde{P}\right)_{x_i},
\end{aligned}$$

where

$$Q(P,\delta) = (1-\delta)\sum_{i=0}^{\infty}\delta^i P^i = (1-\delta)\left(I-\delta P\right)^{-1}. \qquad (6)$$

Analogous to (4), we can define the follwoing distribution induced on $\mathcal{S}$

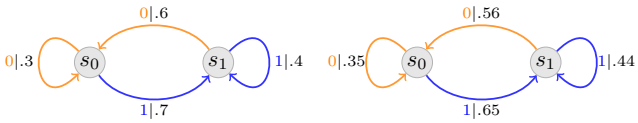
$$\mathbf{p}_i = \frac{\mathbf{p}_{i-1}^T Q(P,\delta)\Gamma_{x_i}}{\left\|\mathbf{p}_{i-1}^T Q(P,\delta)\Gamma_{x_i}\right\|_1}. \qquad (7)$$

Comparing (7) with expressions $\mathbf{p}_i$ in (4), the result follows. ∎

Remark 1. Notice that while the row-stochastic matrix P may not be invertible, $I-\delta P$ is non-singular for all $\delta \in [0,1)$, as the the eigenvalues of P are less than or equal to 1. Moreover, it is clear from (6) that $Q(P,\delta)$ is also a row-stochastic matrix with $\mathbf{p}$ being its eigenvector corresponding to eigenvalue one. We will give a closer look at the eigenvalues of $Q(P,\delta)$ in the next section.

C. Properties of the generalized PFSA

We start by analyzing the eigenspace of the state-to-state transition matrix of $g(\delta)$. Note that it follows from (1) that $P_{g(\delta)} = Q\left(P_g,\delta\right)P_g$.



(a) $\delta = 0$ (b) $\delta = .25$

Fig. 3. On the left: $g_{(.3,.6)}$ in class M2. On the right, $g_{(.3,.6)}(.25)$, with transition probabilities rounded to two decimal places. We can see that deletion only cause the transition probabilities to change, but keep the structure of the machine.

Theorem 2. *Let $\mathbf{p}_g$ be the stationary distribution of strongly connected g. Then the generalized PFSA $g(\delta)$ is also strongly connected with stationary distribution $\mathbf{p}_{g(\delta)} = \mathbf{p}_g$.*

Proof. Let λ be an eigenvalue of P_g. Then $\lambda(1-\delta)(1-\delta\lambda)^{-1}$ is an eigenvalue of $P_{g(\delta)}$. Define $f(\lambda,\delta) = \lambda(1-\delta)(1-\delta\lambda)^{-1}$. Then the result follows from the following observations:

1) For $\lambda = 1$, $f(\lambda,\delta) = 1$ for all $\delta \in [0,1)$, and hence $\lim_{\delta\to 1} f(1,\delta) = 1$.
2) For $\lambda < 1$, $f(\lambda,\delta) < \lambda$ for all $\delta \in [0,1)$, and furthermore, $\lim_{\delta\to 1} f(\lambda,\delta) = 0$. ∎

Then following is an immediate corollary.

Corollary 1. *We have for all $x \in \mathcal{X}$*

$$p_g(x) = p_{g(\delta)}(x).$$

Proof. We have

$$\mathbf{p}_{g(\delta)}^T \widetilde{P}_{g(\delta)} = \mathbf{p}_g^T \widetilde{P}_{g(\delta)} = \mathbf{p}_g^T Q(P_g,\delta)\widetilde{P}_g = \mathbf{p}_g^T \widetilde{P}_g.$$ ∎

A natural question is what happens when $\delta \uparrow 1$. Letting $g(1)$ denote the machine corresponding to $\delta \uparrow 1$, we now show that, quite expectedly, $g(1)$ is a single-state machine.

Theorem 3. *$g(1)$ is a single-state PFSA.*

Proof. First note that the observations given in the proof of Theorem 2 imply that

$$\lim_{\delta\to 1} Q(P_g,\delta) = \mathbf{1}_{|\mathcal{S}|}\mathbf{p}_g^T,$$

and consequently $g(1)$ is a PFSA specified by $\mathbf{1}_{|\mathcal{S}|}\mathbf{p}_g^T\Gamma_{g,x}$ for $x \in \mathcal{X}$.

Suppose $\mathbf{x} = x_1x_2\ldots x_n$ is observed. Following the argument given in section IV-B, we get

$$\begin{aligned} & p_{g(1)}(\mathbf{x}x) \\ = \; & \mathbf{p}^T(\mathbf{1}\mathbf{p}^T\Gamma_{x_1})(\mathbf{1}\mathbf{p}^T\Gamma_{x_2})\cdots(\mathbf{1}\mathbf{p}^T\Gamma_{x_n})\left(\widetilde{P}_{g(1)}\right)_{\cdot,x} \\ = \; & \mathbf{p}^T(\mathbf{1}\mathbf{p}^T\Gamma_{x_1})(\mathbf{1}\mathbf{p}^T\Gamma_{x_2})\cdots(\mathbf{1}\mathbf{p}^T\Gamma_{x_n})\left(\mathbf{1}\mathbf{p}^T\Gamma_x\mathbf{1}\right) \\ = \; & \left(\mathbf{p}^T\mathbf{1}\right)\left(\mathbf{p}^T\Gamma_{x_1}\mathbf{1}\right)\cdots\left(\mathbf{p}^T\Gamma_{x_n}\mathbf{1}\right)\left(\mathbf{p}^T\Gamma_x\mathbf{1}\right), \end{aligned}$$

and hence, by induction, $p_{g(1)}(x|\mathbf{x}) = \mathbf{p}^T\Gamma_x\mathbf{1}$ for all $\mathbf{x}$. Since an i.i.d. process corresponds to a single-state PFSA, we conclude that $g(1)$ is in fact a single-state PFSA. ∎

D. M2 Class of PFSA

We note that $g(\delta)$ of a PFSA g is not necessarily a PFSA. As an example, the Γ-expression of the generalized PFSA $g(.4)$ for g being the PFSA described in Fig. 1 is

$$\Gamma_{g(.4),0} = \begin{pmatrix} .26 & 0 & .14 & 0 \\ .16 & 0 & .44 & 0 \\ .57 & 0 & .08 & 0 \\ .14 & 0 & .40 & 0 \end{pmatrix}, \Gamma_{g(.4),1} = \begin{pmatrix} 0 & .50 & 0 & .10 \\ 0 & .08 & 0 & .32 \\ 0 & .29 & 0 & .06 \\ 0 & .07 & 0 & .39 \end{pmatrix}.$$

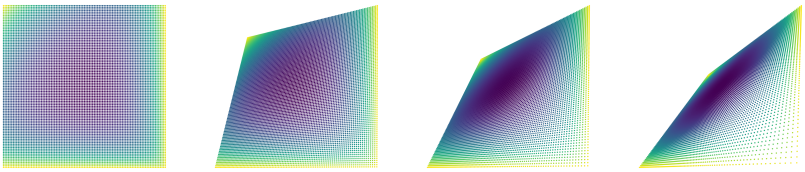
Nevertheless, we introduce M2 a class of PFSAs which is closed under deletion, i.e. $g \in$ M2 implies $g(\delta) \in$ M2 for all $\delta \in [0,1]$. As this class is instrumental in our experimental results, we shall study it in more details.

M2 is the collection of 2-state PFSAs on a binary alphabet: $g = g_{(\mu,\nu)} \in$ M2 with $\mu,\nu \in (0,1)\times(0,1)$ is specified by a quadruple $\left(\mathcal{S},\mathcal{X},\mathsf{T},\mathsf{P}_{(\mu,\nu)}\right)$, where $\mathcal{S} = \{s_0,s_1\}$ $\mathcal{X} = \{0,1\}$, and

$$\Gamma_{g_{(\mu,\nu)},0} = \begin{pmatrix} \mu & 0 \\ \nu & 0 \end{pmatrix}, \quad \Gamma_{g_{(\mu,\nu)},1} = \begin{pmatrix} 0 & 1-\mu \\ 0 & 1-\nu \end{pmatrix}.$$

Fig. 3 illustrates $g_{(.3,.6)}$ and its corresponding $g_{(.3,.6)}(\delta)$, which is obtained from Theorem 1. Since $\Gamma_{g,x,\delta}$ has exactly the same form – containing a single column of non-zero entries for all δ, it is clear that $g_{(.3,.6)}(\delta) \in$ M2.

Since each $g_{(\nu,\mu)}$ is specified by two numbers, we can parametrize M2 by a square in $\mathbb{R}^2$. In Fig. 4, we show the effect of deletion probability on M2 machines. The key observation is that deletion probability drives machines to $\mu = \nu$ line.



(a) $\delta = 0$ (b) $\delta = 0.25$ (c) $\delta = 0.50$ (d) $\delta = 0.75$

Fig. 4. Each dot in (a) represents a $g_{(\mu,\nu)}$ in M2 with μ, ν both ranging from 0.01 to 0.99 and with 0.01 increment. The color of the points is proportional to the KL divergence (defined in Section V-B) of $g_{(.5,.5)}$ to g. The reason that the images are symmetric with respect to the $\mu + \nu = 1$ line is that $g_{(1-\nu,1-\mu)}$ is exactly $g_{(\mu,\nu)}$ with the two states swapped. We can see that while we increase δ, the dots are moving towards the $\mu = \nu$ line which corresponds to the single-state PFSA. The asymmetry in how fast PFSA on each side of the $\mu + \nu = 1$ line converges to single-state PFSA is caused by structural difference between them – machines on the upper side, with $\mu < \nu$, have strong connections between two states, while machines on the lower side, with $\mu > \nu$, have weaker connection between the states.

V. The Convergence of Likelihood

The goal of this section is to lay the theoretical ground for our algorithms for decoding and tamper detecting with PFSAs. In Section V-C, we employ maximum likelihood framework to decode the generating PFSA given the channel output. We show that likelihood is closely related to entropy rate and KL divergence of PFSAs (to be defined and calculated in V-A and V-B).

A. Entropy rate of PFSA

Let g be a PFSA. We define $H_n(g)$ as the following:

$$H_n(g) \coloneqq -\sum_{|\mathbf{x}|=n} p_g(\mathbf{x}) \log p_g(\mathbf{x}).$$

Then the entropy rate of g is defined as

$$H(g) \coloneqq \lim_{n\to\infty} \frac{1}{n} H_n(g).$$

Note that $H(g)$ is in fact the entropy rate of the stochastic process corresponding to g [15]. In the next theorem, we show that the above limit exists and and the entropy rate has a simple closed form.

Theorem 4. *We have*

$$H(g) = \sum_{s\in\mathcal{S}} (\mathbf{p}_g)_s H\left(\left(\widetilde{P}_g\right)_{s,\cdot}\right)$$

Proof. See Appendix VII-A. ∎

It readily follows from the theorem above that the entropy rate for $g_{(\mu,\nu)}$ is

$$H\left(g_{(\mu,\nu)}\right) = \frac{\nu h_b(\mu)}{\bar{\mu}+\nu} + \frac{\bar{\mu} h_b(\nu)}{\bar{\mu}+\nu},$$

where $\bar{a} \coloneqq 1 - a$ and $h_b(a) \coloneqq -a \log a - \bar{a} \log \bar{a}$ is the binary entropy function for any $a \in [0, 1]$.

Next, we show that deletion increases entropy rate, which will be critical for tamper detection purpose.

Theorem 5. *The map $\delta \mapsto H(g_{(\mu,\nu)}(\delta))$ is monotonically increasing when $\mu \neq \nu$.*

Proof. We have

$$\mu(\delta) = \frac{\mu - \delta(\mu-\nu)}{1-\delta(\mu-\nu)}, \quad \nu(\delta) = \frac{\nu}{1-\delta(\mu-\nu)},$$

and

$$H\left(g_{(\mu,\nu)}(\delta)\right) = \frac{\nu}{1-\mu+\nu} h_b\left(\frac{\mu-\delta(\mu-\nu)}{1-\delta(\mu-\nu)}\right) + \frac{1-\mu}{1-\mu+\nu} h_b\left(\frac{\nu}{1-\delta(\mu-\nu)}\right).$$

We can then write

$$\frac{\mathrm{d}}{\mathrm{d}\delta} H(g_{(\mu,\nu)}(\delta)) = \frac{\alpha\bar{\mu}\nu}{(1-\alpha\delta)^2\bar{\alpha}} \log \frac{(\mu-\delta\alpha)(\bar{\nu}-\delta\alpha)}{\bar{\mu}\nu},$$

where $\alpha = \mu - \nu$. It's straightforward to check that the derivative is always positive when $\mu \neq \nu$. ∎

B. KL divergence of two PFSAs

Let $g_1, g_2 \in$ M2. The n-th order KL divergence between g_1 and g_2 is the KL divergence on the space of length-n sequences, i.e.

$$D_n\left(g_1 \| g_2\right) = \sum_{|\mathbf{x}|=n} p_{g_1}(\mathbf{x}) \log \frac{p_{g_1}(\mathbf{x})}{p_{g_2}(\mathbf{x})}.$$

Analogous to entropy rate, we can define the KL divergence between g_1 and g_1 as

$$D_{\mathrm{KL}}\left(g_1 \| g_2\right) \coloneqq \lim_{n\to\infty} \frac{1}{n} D_n\left(g_1 \| g_2\right).$$

We show in Theorem 6 below shows that the limit exists and also derived a closed form for the KL divergence between two PFSAs. But before we can state the theorem, we need to introduce a very useful construction on two PFSAs, called *synchronous composition*.

Definition 4 (synchronous composition)**.** *Let $g_1 = (\mathcal{S}, \mathcal{X}, \mathsf{T}_1, \mathsf{P}_1)$ and $g_2 = (\mathcal{T}, \mathcal{X}, \mathsf{T}_2, \mathsf{P}_2)$ be two PFSAs with the same alphabet and let $g_c^*\left(g_1 \| g_2\right)$ be the probabilistic automata specified by the quadruple $(\mathcal{S}_c, \mathcal{X}, \mathsf{T}_c, \mathsf{P}_c)$ where*

$$\mathcal{S}_c = \mathcal{S}_1 \times \mathcal{T} = \{(s,t)\}_{s\in\mathcal{S}_1, t\in\mathcal{T}}$$

is the Cartesian product of $\mathcal{S}$ and $\mathcal{T}$, and

$$\mathsf{T}_c((s,t),x) = \left(\mathsf{T}_1(s,x), \mathsf{T}_2(t,x)\right),$$
$$\mathsf{P}_c((s,t),x) = \mathsf{P}_1(s,x),$$

for all $s \in \mathcal{S}$, $t \in \mathcal{S}$, and $x \in \mathcal{X}$. Then the synchronous composition $g_c\left(g_1 \| g_2\right)$ is defined to be any absorbing strongly connected component of $g_c^*\left(g_1 \| g_2\right)$, i.e. strongly connected component without any out-going edges.

It is not clear that there is only one absorbing strongly connected component in $g_{\mathrm{c}}^*\left(g_1 \| g_2\right)$. However, as proved in Theorem 8 in Appendix VII-B, $g_{\mathrm{c}}\left(g_1 \| g_2\right)$ is *equivalent* to g_1 irrespective of the choice of absorbing strongly connected component, i.e., $p_{g_{\mathrm{c}}}(\mathbf{x}) = p_{g_1}(\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}^*$.

In Figs. 6, 7, 8, and 9, we provide examples of synchronous compositions for several g_1 and g_2 which shed light on the fact that the synchronous composition of two strongly connected PFSA might not be strongly connected.

Theorem 6. *Let $g_c = g_c\left(g_1 \| g_2\right)$ and $\mathbf{p}_{g_c}$ be the stationary distribution of g_c. Then we have*

$$\lim_{n\to\infty} \frac{1}{n} D_n\left(p_{g_1}^n \| p_{g_1}^n\right) = \sum_{s\in\mathcal{S}, t\in\mathcal{T}} \left(\mathbf{p}_{g_c}\right)_{(s,t)} D_{KL}\left(\left(\widetilde{P}_{g_1}\right)_{s,\cdot} \middle\| \left(\widetilde{P}_{g_2}\right)_{t,\cdot}\right).$$

Proof. See Appendix VII-B. ∎

In light of this theorem, one can easily show

$$D_{\mathrm{KL}}\left(g_1 \| g_2\right) = \frac{\nu_1 D_{\mathrm{KL}}\left(\mu_1 \| \mu_2\right)}{\bar{\mu}_1 + \nu_1} + \frac{\bar{\mu}_1 D_{\mathrm{KL}}\left(\nu_1 \| \nu_2\right)}{\bar{\mu}_1 + \nu_1}.$$

C. Convergence of log likelihood

According to Shannon-McMillan-Breiman Theorem [15, Theorem 16.8.1], we have $-\frac{1}{n}\log p_g(\mathbf{x}) \to H(g)$ for any sequence $\mathbf{x} \leftarrow g$. A natural question is that what the log-likelihood converges to if $\mathbf{x}$ is generated by a different machine. The following theorem states that the log-likelihood converges to entropy of generating machine plus the KL divergence which accounts for the mismatch.

Theorem 7. *For any $\mathbf{x}^n \leftarrow g \in M2$, we have with probability one*

$$-\frac{1}{n}\sum_{i=1}^{n} \log p_{g'}\left(x_i | \mathbf{x}^{i-1}\right) \to H(g) + D_{KL}\left(g \| g'\right),$$

for any PFSA $g' \in M2$.

Proof. First note that

$$-\frac{1}{n}\sum_{i=1}^{n}\log p_{g'}(x_i|\mathbf{x}^{i-1}) = -\frac{1}{n}\log p_g(\mathbf{x}) + \frac{1}{n}\sum_{i=1}^{n}\log\frac{p_g(x_i|\mathbf{x}^{i-1})}{p_{g'}(x_i|\mathbf{x}^{i-1})}. \tag{8}$$

Clearly, the first term in the above sum converges to $H(g)$. To show the convergence of the second term, let $Z_i = \log\frac{p_g\left(x_i|\mathbf{x}^{i-1}\right)}{p_{g'}(x_i|\mathbf{x}^{i-1})}$. Notice that for any PFSA g in M2 and for $1 \le i \le n$, $\mathbf{p}_g(\mathbf{x}^i)$ equals $[1,0]$ for all $\mathbf{x}^i$ with $x_i = 0$, and to $[0,1]$ for all $\mathbf{x}^i$ with $x_i = 1$, and hence the process $\{Z_i\}_{i=1}^n$ is a Markov process. Let $\mathcal{Z}^0$ and $\mathcal{Z}^1$ denote the set of indices i such that $x_{i-1} = 0$ and $x_{i-1} = 1$, respectively. Then we have

$$\frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{1}{n}\sum_{i\in\mathcal{Z}_0} Z_i + \frac{1}{n}\sum_{i\in\mathcal{Z}_1} Z_i. \tag{9}$$

It is straightforward to show that for all $i \in \mathcal{Z}_0$

$$Z_i = 1_{\{x_i=0\}}\log\frac{\mu_g}{\mu_{g'}} + 1_{\{x_i=1\}}\log\frac{\bar{\mu}_g}{\bar{\mu}_{g'}},$$

and for all $i \in \mathcal{Z}_1$

$$Z_i = 1_{\{x_i=0\}}\log\frac{\nu_g}{\nu_{g'}} + 1_{\{x_i=1\}}\log\frac{\bar{\nu}_g}{\bar{\nu}_{g'}}.$$

It follows from (9) that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} Z_i \quad = \quad & \frac{1}{n}\left(\log\frac{\mu_g}{\mu_{g'}}\right)\sum_{i=1}^{n} 1_{\{x_{i-1}=0,x_i=0\}} + \frac{1}{n}\left(\log\frac{\bar{\mu}_g}{\bar{\mu}_{g'}}\right)\sum_{i=1}^{n} 1_{\{x_{i-1}=0,x_i=1\}} + \frac{1}{n}\left(\log\frac{\nu_g}{\nu_{g'}}\right)\sum_{i=1}^{n} 1_{\{x_{i-1}=1,x_i=0\}} \\ & + \frac{1}{n}\left(\log\frac{\bar{\nu}_g}{\bar{\nu}_{g'}}\right)\sum_{i=1}^{n} 1_{\{x_{i-1}=1,x_i=1\}} \\ \xrightarrow{n\to\infty} \quad & \mathbf{p}_g(0)\left(\mu_g\log\frac{\mu_g}{\mu_{g'}} + \bar{\mu}_g\log\frac{\bar{\mu}_g}{\bar{\mu}_{g'}}\right) + \mathbf{p}_g(1)\left(\nu_g\log\frac{\nu_g}{\nu_{g'}} + \bar{\nu}_g\log\frac{\bar{\nu}_g}{\bar{\nu}_{g'}}\right). \quad \blacksquare\end{aligned}$$

For ease of presentation, we define

$$L\left(g', \mathbf{x}^n \leftarrow g\right) \coloneqq -\frac{1}{n}\sum_{i=1}^{n}\log p_{g'}(x_i|\mathbf{x}^{i-1}).$$

When the generating machine g is not known, we use $L\left(g', \mathbf{x}^n\right)$ to identify likelihood of g' generating x.

VI. Algorithm and Simulation

A. Decoding

In this and the following section, we assume that we have a set of PFSAs $\mathcal{G} = \left\{g_1, \ldots, g_{|\mathcal{M}|}\right\}$, with $g_i \in$ M2 for all i. We will briefly discuss heuristics on how to generate a set of PFSAs that are good for tamper detecting and decoding in SectionVI-C.

We saw in Theorem 7 that

$$L\left(g_j(\delta), \mathbf{x}^n \leftarrow g_i(\delta)\right) \to H(g_i(\delta)) + D_{\mathrm{KL}}\left(g_i(\delta) \| \, g_j(\delta)\right), \tag{10}$$

which motivates the following definition for the decoding function in Fig. 2

$$\psi(\mathbf{x}) = \arg\min_{m\in\mathcal{M}} L\left(g_m(\delta), \mathbf{x}^n\right).$$

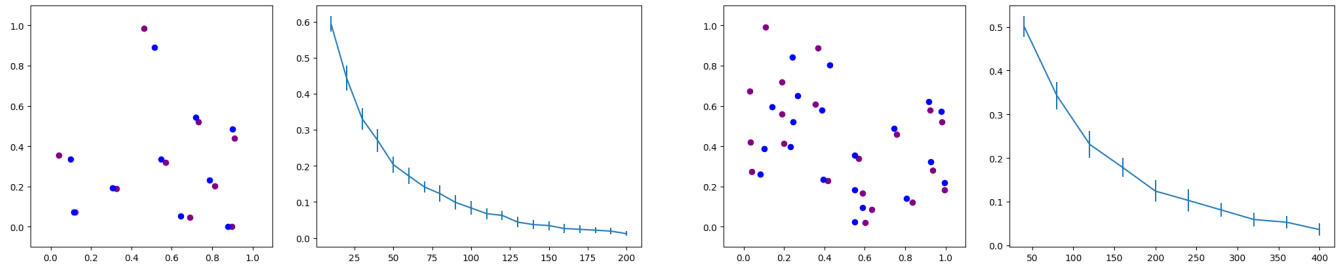
We apply this decoding strategy in Fig. 5 when $\delta = .2$ and two different message sets with $|\mathcal{M}| = 10$ or $|\mathcal{M}| = 20$.

B. Tamper detecting

We assume that active eavesdropper tampers the channel in such a way that $\delta' - \delta > \eta$ with some $\eta \ge 0$. Following Theorems 5 and 7, we get

$$\begin{aligned} L\left(g_j(\delta), \mathbf{x} \leftarrow g_i(\delta')\right) &\to H(g_i(\delta')) + D_{\mathrm{KL}}\left(g_i(\delta') \| \, g_j(\delta)\right) \\ &\ge H(g_i(\delta)), \end{aligned} \tag{11}$$

where the inequality is due to Theorem 5. Hence, tampering the channel results in an increase in the likelihood. This leads to our temper detecting procedure detailed in Algorithm 1.



(a) (b) (c) (d)

Fig. 5. (a) and (c) shows 10 PFSAs and 20 PFSAs in the parameter space, respectively, with purple dots for the g_m's and blue dots for $g_m(.2)$'s. Error rates for input sequences of length 10 to 200 for the 10 messages, and for input sequences of length 40, 400 for the 20 messages are showed in (b) and (d), respectively. The results are averaged over 20 and 10 re-runs.

Algorithm 1: Tampering detection

input : $\{g_m\}_{m\in\mathcal{M}}$, $\mathbf{x}_1, \ldots, \mathbf{x}_k$, δ, η, ε
output: T with $T = 0$ if no tampering, 1 if otherwise
$H_0 = (H(g_m(\delta)))_{m\in\mathcal{M}}$;
$H_1 = (H(g_m(\delta+\eta)))_{m\in\mathcal{M}}$;
$D = H_1 - H_0$;
$v = 0$; /* the weighted vote */
for $i = 1, \ldots, k$ **do**
  $d = \arg\min_{m\in\mathcal{M}} L(g_m(\delta), \mathbf{y}_i)$;
  $e = L(g_d(\delta), \mathbf{y}_i)$;
  if $e - H(g_d(\delta)) > \varepsilon \cdot D[d]$ **then**
    $v = v + 1 \cdot D[d]$;
  end
end
$S = \sum_{m\in\mathcal{M}} D[m]$;
if $v/(S \cdot k) > 0.5$ **then**
  return $T = 1$;
else
  return $T = 0$;
end

C. Generate machines with good separation

For fixed number of messages, we need to choose a set of M2 PFSAs with the best decoding and tamper detection performance. It is important to indicate that (1) decoding error will be significantly lowered by increasing $D(g_i\|g_j)$ according to (10), and (2) the tampering detection error will be improved by making sure $|H(g(\delta)) - H(g(\delta'))|$ is large for $\delta' - \delta \geq \eta$, according to (11). However, there is a trade-off here – to increase pairwise KL divergence, we want the machines to be spread more evenly in the parameter space while, according to Theorem 5, to increase $H(g(\delta')) - H(g(\delta)$, we need the machines to stay away from being single-state, i.e. away from the $\mu = \nu$ line.

Here, we describe briefly how we design $\mathcal{G}$ for experiment in Fig. 5. As a naive way, we start off with $|\mathcal{M}|$ randomly generated μ's in $(0, 1)$, and for each of them we generate ν in the following way: if $\mu > .5$, then we choose a ν randomly in $(0, \mu - .2)$, and if $\mu \leq .5$, in $(\mu + .2, 1.)$. Then, we use a hill-climbing algorithm to maximize minimum pairwise averaged KL divergence, $.5\,(D_{\mathrm{KL}}\,(g_1\|\,g_2\,) + D_{\mathrm{KL}}\,(g_2\|\,g_1\,))$, between all pair machines. Let σ be step size, for a pair $g_{(\mu_1,\nu_1)}$ and $g_{(\mu_2,\nu_2)}$ with minimum averaged KL divergence, we search the eight neighboring points, $(\mu_i \pm \sigma, \nu_i)$ and $(\mu, \nu_i \pm \sigma)$, $i = 1, 2$, for improvement. We exit the search when there is no improvement to be found.

VII. Conclusion and Future Work

In this paper, we developed a new information-theoretic coding scheme for information transfer over a public deletion channel, subject to an active eavesdropper (aka jammer). Our coding scheme is based on probabilistic finite-state automata (PFSA) and is proved to have (1) semi-universal property, in a sense that codebook need not be available at the decoder, (2) small error probability when decoding messages, and (3) tamper-free property, which alarms the decoder about possible tampering of the channel. To the best of our knowledge, exploiting PFSA's in a secure and reliable information-theoretic

	$\varepsilon = 0$			$\varepsilon = 0.05$			$\varepsilon = 0.10$		
50	.16	.12	.28	.26	.08	.34	.24	.18	.32
100	0	.26	.26	.06	.20	.26	.08	.08	.16
150	0	.16	.16	.02	.22	.14	0	.08	.08
200	0	.18	.18	0	.20	.20	0	.06	.06
	$\varepsilon = 0.15$			$\varepsilon = 0.20$			$\varepsilon = 0.25$		
50	.36	.02	.38	.32	.02	.34	.26	.10	.36
100	.08	.08	.16	.08	.02	.10	.14	.04	.18
150	0	.08	.08	.02	.02	.04	.02	.02	.04
200	0	0	0	0	.04	.04	0	.02	.02

Table I. The table above records the error rates of tamper detection algorithm for sending 10 messages through a channel with deletion probability $\delta = .2$. We generate 50 test sets containing $k = 200$ sequences, with 20 for each message. We assign randomly whether a particular test set will be tampered or not. For simplicity, if a test set is tampered it will have a fixed deletion probability $\delta = .3$. We run the algorithm for input sequence of length $50, 100, 150$, and 200, and for $\varepsilon = 0, .05, .10, .15, .20, .25$. For each block, the first column is the rate of failing to detect a tampering, and the second column is the rate of false alarm of a tampering, and the last column is the sum of two error rates. We can see that with increased cutoff value ε, we have significantly fewer false alarms without too much increase in the rate of failing to detect a true tampering.

communication model is very new, yet very insightful. Promising results in both theoretical and experimental aspects of this work lead to several research directions:

- To have an analytically better analysis of error probability, the convergence rate of likelihood in Theorem7 for general PFSA is needed.
- We admit that the space of M2 is too small to have simultaneous vanishing error probability (with small n) in message decoding and tamper detecting. To go beyond M2, we need to find an analytic way to compute entropy rate and KL divergence for generalized PFSA.

REFERENCES

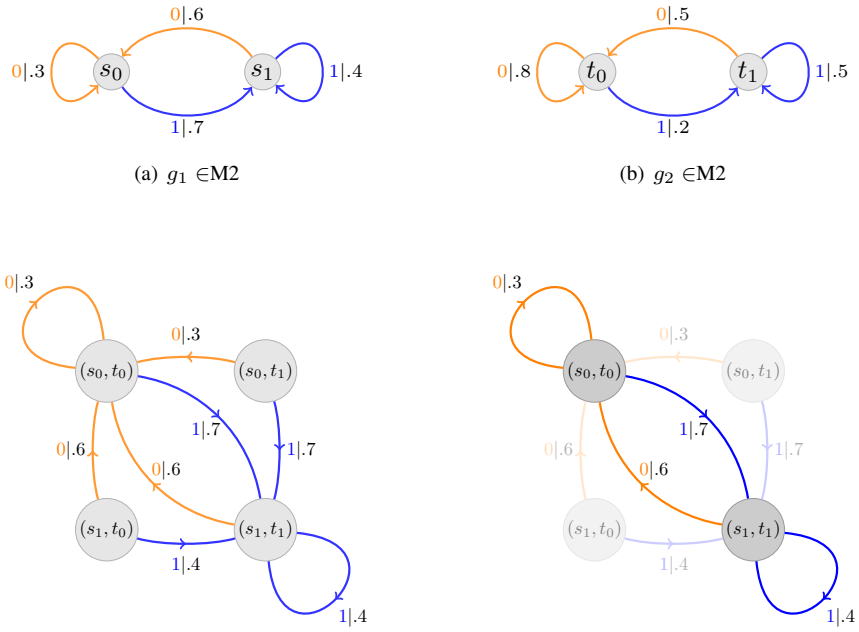
[1] M. Mitzenmacher, "A survey of results for deletion channels and related synchronization channels," *Probability Surveys*, vol. 6, pp. 1–33, 2009.
[2] R. L. Dobrushin, "Shannon's theorems for channels with synchronization errors," *Problems Inform. Transmission*, vol. 3, no. 4, pp. 11–26, Oct. 1967.
[3] A. Kirsch and E. Drinea, "Directly lower bounding the information capacity for channels with i.i.d. deletions and duplications," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 86–102, Jan 2010.
[4] M. Mitzenmacher and E. Drinea, "A simple lower bound for the capacity of the deletion channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4657–4660, Oct. 2006.
[5] H. Yamamoto, "A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers," *IEEE Trans. Inf. Theory*, vol. 29, no. 6, pp. 918–923, Nov. 1983.
[6] D. Gündüz, E. Erkip, and H. Poor, "Secure lossless compression with side information," in *Proc. IEEE Information Theory Workshop*, May 2008, pp. 169–173.
[7] E. Ekrem and S. Ulukus, "Secure lossy source coding with side information," in *Proc. Annual Allerton Conference on Communication, Control, and Computing*, Sept. 2011, pp. 1098–1105.
[8] Y.-H. Kim, A. Sutivong, and T. Cover, "State amplification," *IEEE Trans. Inf. Theory,*, vol. 54, no. 5, pp. 1850–1859, May 2008.
[9] K. Kittichokechai, Y. K. Chia, T. J. Oechtering, M. Skoglund, and T. Weissman, "Secure source coding with a public helper," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 3930–3949, July 2016.
[10] Y. Kaspi and N. Merhav, "Zero-delay and causal secure source coding," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 6238–6250, Nov 2015.
[11] J. Villard and P. Piantanida, "Secure multiterminal source coding with side information at the eavesdropper," *IEEE Trans. Inf Theory*, vol. 59, no. 6, pp. 3668–3692, June 2013.
[12] S. Asoodeh, F. Alajaji, and T. Linder, "Lossless secure source coding: Yamamoto's setting," in *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2015, pp. 1032–1037.
[13] V. Misra and T. Weissman, "Unsupervised learning and universal communication," in *IEEE Inter. Symp. Inf. Theory*, July 2013, pp. 261–265.
[14] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 2, 1948.
[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.

APPENDIX

A. *Proof for Theorem 4*

Following the standard notation in information theory, we use X^n to denote a random vector $(X_1, \ldots, X_n)$ generated from a PFSA g and $H(X^n)$ to denote the entropy its entropy, that is $H(X^n) = H_n(g)$. We can similarly define the conditional entropy $H(X_n|X^{n-1})$. It is shown in [15] that $\lim_{n\to\infty} \frac{1}{n} H(X^n) = \lim_{n\to\infty} H\left(X_n|X^{n-1}\right)$ for any stationary processes $\{X_n\}_{n=1}^{\infty}$. In order to compute the entropy rate, we can therefore focus on the latter limit. Let $S \sim \mathbf{p}$ denote a random variable indicating the initial state of the PFSA. We have

$$\begin{aligned} H\left(X_n|X^{n-1}\right) &= H\left(X^n\right) - H\left(X^{n-1}\right) \\ &= \left[H\left(X^n, S\right) - H\left(S|X^n\right)\right] - \left[H\left(X^{n-1}, S\right) - H\left(S|X^{n-1}\right)\right] \\ &= \left[H\left(X^n, S\right) - H\left(X^{n-1}, S\right)\right] + \left[H\left(S|X^{n-1}\right) - H\left(S|X^n\right)\right] \end{aligned}$$



(a) g_1 ∈M2

(b) g_2 ∈M2

(c) $g_c^*(g_1 \| g_2)$

(d) strongly connected component of $g_c^*(g_1 \| g_2)$

Fig. 6. The example above shows that the g_c^* of two strongly connected PFSAs may *not* remain strongly connected. We can see that in this case, $g_c(g_1 \| g_2)$ is equal to g_1.

$$= \left[H\left(X^n|S\right) + H(S) - H\left(X^{n-1}|S\right) - H(S)\right] + \left[H\left(S|X^{n-1}\right) - H\left(S|X^n\right)\right]$$
$$= \underbrace{H\left(x_n|S, X^{n-1}\right)}_{=:A_n} + \underbrace{\left[H\left(S|X^{n-1}\right) - H\left(S|X^n\right)\right]}_{=:B_n}.$$

Note that for any $N \geq 1$

$$\sum_{n=1}^{N} B_n = \sum_{n=1}^{N} H\left(S|X^{n-1}\right) - H\left(S|X^n\right) = H(S) - H\left(S|X^N\right) \leq H(S) = H\left(\mathbf{p}\right).$$

Since B_n is nonnegative for each n and $\sum_{n=1}^{N} B_n$ is bounded from above, it follows that $\lim_{n\to\infty} B_n = 0$. It remains to analyze A_n. Notice that the state at time n is a deterministic function of S and X^{n-1} (that is $\mathsf{T}(s, X^{n-1})$) and hence we can write

$$H\left(X_n|S, X^{n-1}\right) = \sum_{s' \in \mathcal{S}} H\left(\widetilde{P}_{s',\cdot}\right) \Pr\left\{\mathsf{T}(S, X^{n-1}) = s'\right\}.$$

By induction, we have for any $s' \in \mathcal{S}$

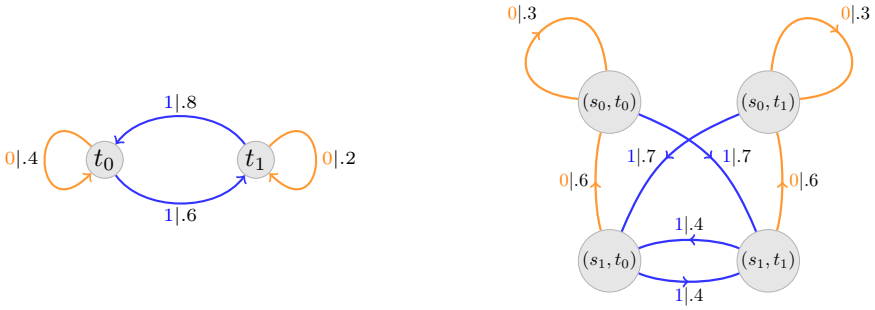
$$\Pr\left\{\mathsf{T}(S, X^{n-1}) = s'\right\} = \sum_{x \in \mathcal{X}} \sum_{s'' \in \mathcal{S}} \Pr\left\{\mathsf{T}(s, X^{n-2}) = s''\right\} \left(\Gamma_x\right)_{s'',s'}$$
$$= \sum_{s'' \in \mathcal{S}} \Pr\left\{\mathsf{T}(s, X^{n-2}) = s''\right\} P_{s'',s'},$$

and hence

$$\left(\Pr\left\{\mathsf{T}(s, X^{n-1}) = s\right\}\right)_{s \in \mathcal{S}} = \left(\Pr\left\{\mathsf{T}(s, X^{n-2}) = s\right\}\right)_{s \in \mathcal{S}} P = \cdots = \mathbf{p}P^{n-1} = \mathbf{p}.$$

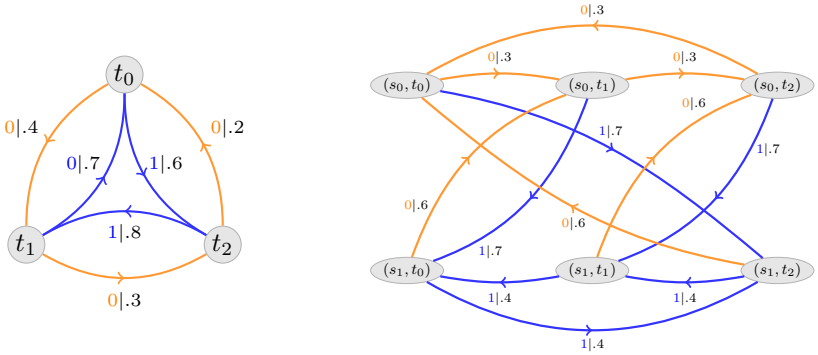
B. Proof for Theorem 6

Before we can prove Theorem 6, we first study synchronous compositions in more detail. Specifically, we shall show that $\mathbf{p}_{g_c}(\mathbf{x})$ is independent of the choice of absorbing strongly connected component in $g_c^*(g_1 \| g_2)$. Essentially, $g_c(g_1 \| g_2)$ is equivalent (to be defined later) to g_1, which is key to the usage of synchronous composition in the proof of Theorem 6.



(a) g_3 (b) $g_c\,(g_1 \| \, g_3\,)$

Fig. 7. $g_c\,(g_1 \| \, g_3\,)$ is strongly connected. The stationary distribution of $g_c\,(g_1 \| \, g_3\,)$ is $(.231, .231, .269, .269)$, while the stationary distribution of g_1 is $(.462, .538)$, both rounded to 3 decimal places.



(a) g_4 (b) $g_c\,(g_1 \| \, g_4\,)$

Fig. 8. $g_c\,(g_1 \| \, g_4\,)$ is strongly connected. The stationary distribution of $g_c\,(g_1 \| \, g_4\,) = (.154, .154, .154, .179, .179, .179)$, while the stationary distribution of g_1 is $(.462, .538)$, both rounded to 3 decimal places.

Definition 5. *Let $g_1 = (\mathcal{S}, \mathcal{X}, \mathsf{T}_1, \mathsf{P}_1)$ and $g_2 = (\mathcal{T}, \mathcal{X}, \mathsf{T}_2, \mathsf{P}_2)$ be two PFSAs with the same alphabet and let $g_c\,(g_1 \| \, g_2\,)$ be the synchronous composition of g_1 and g_2. Suppose that the state space of $g_c\,(g_1 \| \, g_2\,)$ is $\mathcal{U} \subset \mathcal{S} \times \mathcal{T}$. We then define $\mathcal{T}_s = \{t \in \mathcal{T} : (s, t) \in \mathcal{U}\}$.*

We provided several examples of synchronous compositions in Figs. 6 to 9. We note that, in Fig. 7 and 8, the compositions g_c^* are naturally strongly connected, while those in Fig. 6 and 9 are not. For $g_c\,(g_1 \| \, g_2\,)$ in Fig. 6, we have $\mathcal{T}_{s_0} = \{t_0\}$ and $\mathcal{T}_{s_1} = \{t_1\}$, and for $g_c\,(g_5 \| \, g_2\,)$ in Fig. 9, we have $\mathcal{T}_{s_0} = \{t_0\}$, $\mathcal{T}_{s_1} = \{t_1\}$, $\mathcal{T}_{s_2} = \{t_0\}$, and $\mathcal{T}_{s_3} = \{t_1\}$.

Proposition 1. *Let $g_c = g_c\,(g_1 \| \, g_2\,)$ be any absorbing strongly connected component of $g_c^*\,(g_1 \| \, g_2\,)$ and let $\mathbf{p}_{g_c}$ be its stationary distribution. Then we have $\sum_{t \in \mathcal{T}_s} (\mathbf{p}_{g_c})_{(s,t)} = (\mathbf{p}_{g_1})_s$.*

Proof. For any fixed initial state (s, t) and any sequence of symbols $\mathbf{x}^n \in \mathcal{X}^n$, consider the sequence of states of the synchronous composition

$$(s, t), (\mathsf{T}_1\,(s, x_1)\,, \mathsf{T}_2\,(t, x_1))\,, \ldots, (\mathsf{T}_1\,(s, \mathbf{x}^n)\,, \mathsf{T}_2\,(t, \mathbf{x}^n))\,.$$

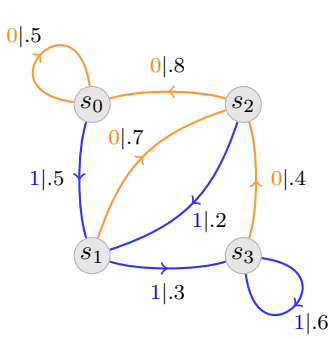
Let $n_{s',t'}$ be the number of indices $i = 1, \ldots, n$ such that $\left(\mathsf{T}_1\left(s, \mathbf{x}^i\right), \mathsf{T}_2\left(t, \mathbf{x}^i\right)\right) = (s', t')$. Since the associated stochastic process on states induced by g_c is stationary and ergodic, we have $\frac{n_{s',t'}}{n} \to (\mathbf{p}_{g_c})_{(s',t')}$ as $n \to \infty$ in probability. Consequently,

$$\sum_{t' \in \mathcal{T}_s} \frac{n_{s',t'}}{n} \to \sum_{t' \in \mathcal{T}_s} (\mathbf{p}_{g_c})_{(s',t')}\,.$$

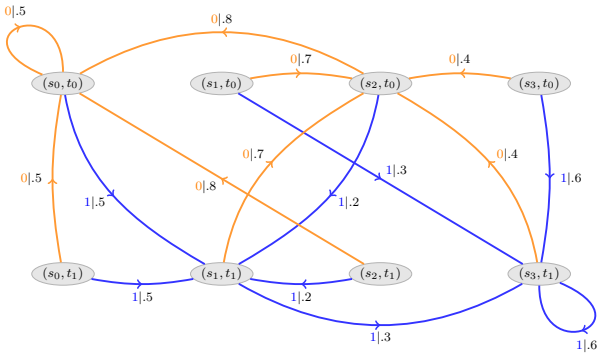
Noticing that the left-hand side converges to $(\mathbf{p}_{g_1})_s$, we obtain the result. ∎

Figs. 7 and 8 provide examples of the proposition above.

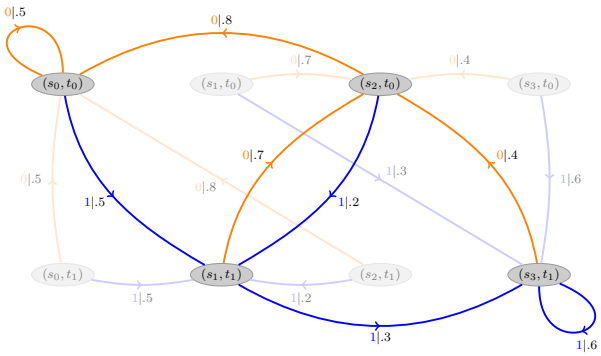
Theorem 8. *Let $g_c = g_c\,(g_1 \| \, g_2\,)$ be any absorbing strongly connected component of $g_c^*\,(g_1 \| \, g_2\,)$. Then we have g_c is equivalent to g_1, in the sense that $p_{g_c}\,(\mathbf{x}) = p_{g_1}\,(\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}^*$.*



(a) g_5



(b) $g_{\mathrm{c}}^{*}\left(g_5 \| \, g_2\right)$



(c) strongly connected component of $g_{\mathrm{c}}^{*}\left(g_5 \| \, g_2\right)$

Fig. 9. $g_{\mathrm{c}}\left(g_5 \| \, g_2\right)$ is the strongly connected component of $g_{\mathrm{c}}^{*}\left(g_5 \| \, g_2\right)$ and it is equal to g_5.

Proof. We first show

$$\sum_{t \in \mathcal{T}_s} p_{g_{\mathrm{c}}}\left(\mathbf{x} | (s,t)\right) \left(\mathbf{p}_{g_{\mathrm{c}}}\right)_{(s,t)} = p_{g_1}\left(\mathbf{x}|s\right) \left(\mathbf{p}_{g_1}\right)_s \tag{12}$$

by induction on the length of $\mathbf{x}$. We first note that the base case in which $\mathbf{x}$ is the empty sequence is given by Proposition 1. Now assume that (12) holds for $|\mathbf{x}| = n$. Follow the notation as in Definition 5, we have for sequence $\mathbf{x}x$

$$\begin{aligned}
\sum_{t \in \mathcal{T}_s} p_{g_{\mathrm{c}}}\left(\mathbf{x}x | (s,t)\right) \left(\mathbf{p}_{g_{\mathrm{c}}}\right)_{(s,t)} &= \sum_{t \in \mathcal{T}_s} p_{g_{\mathrm{c}}}\left(\mathbf{x} | (s,t)\right) p_{g_{\mathrm{c}}}\left(x | \mathbf{x}, (s,t)\right) \left(\mathbf{p}_{g_{\mathrm{c}}}\right)_{(s,t)} \\
&= \sum_{t \in \mathcal{T}_s} p_{g_{\mathrm{c}}}\left(\mathbf{x} | (s,t)\right) \mathsf{P}_1\left(\mathsf{T}_1(s,\mathbf{x}), x\right) \left(\mathbf{p}_{g_{\mathrm{c}}}\right)_{(s,t)} \\
&= \left(\sum_{t \in \mathcal{T}_s} p_{g_{\mathrm{c}}}\left(\mathbf{x} | (s,t)\right) \left(\mathbf{p}_{g_{\mathrm{c}}}\right)_{(s,t)} \right) \mathsf{P}_1\left(\mathsf{T}_1(s,\mathbf{x}), x\right) \\
&\overset{(a)}{=} p_{g_1}\left(\mathbf{x}|s\right) \left(\mathbf{p}_{g_1}\right)_s \mathsf{P}_1\left(\mathsf{T}_1(s,\mathbf{x}), x\right) \\
&= p_{g_1}\left(\mathbf{x}x|s\right) \left(\mathbf{p}_{g_1}\right)_s ,
\end{aligned}$$

where equality in (a) follows from the induction hypothesis. Now we can write

$$p_{g_{\mathrm{c}}}\left(\mathbf{x}\right) = \sum_{s \in \mathcal{S}} \sum_{t \in \mathcal{T}_s} p_{g_{\mathrm{c}}}\left(\mathbf{x} | (s,t)\right) \left(\mathbf{p}_{g_{\mathrm{c}}}\right)_{(s,t)} = \sum_{s \in \mathcal{S}} p_{g_1}\left(\mathbf{x}|s\right) \left(\mathbf{p}_{g_1}\right)_s = p_{g_1}\left(\mathbf{x}\right),$$

from which the result follows. ∎

Proof for Theorem 6. We use the same notation as in Appendix VII-A. We start the proof by defining two distributions on the Cartesian product $\mathcal{S} \times \mathcal{T} \times \mathcal{X}^n$. Let

$$g_{12} \coloneqq g_{\mathrm{c}}\left(g_1 \| \, g_2\right), \quad g_{21} \coloneqq g_{\mathrm{c}}\left(g_2 \| \, g_1\right),$$

and $\mathbf{p}_{12}$ and $\mathbf{p}_{21}$ be the stationary distributions of g_{12} and g_{21}, respectively. Here we make sure that we choose the same absorbing strongly connected component for both compositions. We notice that g_{12} and g_{21} induce two distributions p_{12}, and p_{21} on $\mathcal{S}\times\mathcal{T}\times\mathcal{X}^n$ given by $p_{12}(s,t,\mathbf{x}^{n-1}) = p_{12}(s,t)p_{12}(\mathbf{x}^{n-1}|s,t)$ and $p_{21}(s,t,\mathbf{x}^{n-1}) = p_{21}(s,t)p_{21}(\mathbf{x}^{n-1}|s,t)$ where

$$\begin{aligned}
p_{12}(s,t) &= (\mathbf{p}_{12})_{(s,t)}\,, \quad p_{21}(s,t) = (\mathbf{p}_{21})_{(s,t)}\,,\\
p_{12}\left(\mathbf{x}^n|s,t\right) &= p_{g_1}\left(\mathbf{x}^n|s\right) = \prod_{i=1}^{n} \mathsf{P}_1\left(\mathsf{T}_1\left(s,\mathbf{x}^{i-1}\right),x_i\right),\\
p_{21}\left(\mathbf{x}^n|s,t\right) &= p_{g_2}\left(\mathbf{x}^n|t\right) = \prod_{i=1}^{n} \mathsf{P}_2\left(\mathsf{T}_2\left(t,\mathbf{x}^{i-1}\right),x_i\right).
\end{aligned}$$

Letting $p_{12}(\mathbf{x}^n)$ ($p_{12}(\mathbf{x}^{n-1})$) be the marginal of p_{12} over $\mathcal{X}^n$ (resp. $\mathcal{X}^{n-1}$), we can write using the chain rule of KL divergence (see e.g., [15, Theorem 2.5.3]) that

$$\begin{aligned}
&D_{\mathrm{KL}}\left(p_{12}\left(X^n\right)\|\,p_{21}\left(X^n\right)\right) - D_{\mathrm{KL}}\left(p_{12}\left(X^{n-1}\right)\big\|\,p_{21}\left(X^{n-1}\right)\right)\\
&= \left[D_{\mathrm{KL}}\left(p_{12}\left(S,T,X^n\right)\|\,p_{21}\left(S,T,X^n\right)\right) - D_{\mathrm{KL}}\left(p_{12}\left(S,T|X^n\right)\|\,p_{21}\left(S,T|X^n\right)\right)\right]\\
&\quad - \left[D_{\mathrm{KL}}\left(p_{12}\left(S,T,X^{n-1}\right)\big\|\,p_{21}\left(S,T,X^{n-1}\right)\right) - D_{\mathrm{KL}}\left(p_{12}\left(S,T|X^{n-1}\right)\big\|\,p_{21}\left(S,T|X^{n-1}\right)\right)\right]\\
&= \left[D_{\mathrm{KL}}\left(p_{12}\left(S,T,X^n\right)\|\,p_{21}\left(S,T,X^n\right)\right) - D_{\mathrm{KL}}\left(p_{12}\left(S,T,X^{n-1}\right)\big\|\,p_{21}\left(S,T,X^{n-1}\right)\right)\right]\\
&\quad - \left[D_{\mathrm{KL}}\left(p_{12}\left(S,T|X^n\right)\|\,p_{21}\left(S,T|X^n\right)\right) - D_{\mathrm{KL}}\left(p_{12}\left(S,T|X^{n-1}\right)\big\|\,p_{21}\left(S,T|X^{n-1}\right)\right)\right]\\
&= \underbrace{D_{\mathrm{KL}}\left(p_{12}\left(X_n\left|S,T,X^{n-1}\right.\right)\big\|\,p_{21}\left(X_n\left|S,T,X^{n-1}\right.\right)\right)}_{=:C_n}\\
&\quad - \left[\underbrace{D_{\mathrm{KL}}\left(p_{12}\left(S,T|X^n\right)\|\,p_{21}\left(S,T|X^n\right)\right)}_{=:D_n} - \underbrace{D_{\mathrm{KL}}\left(p_{12}\left(S,T|X^{n-1}\right)\big\|\,p_{21}\left(S,T|X^{n-1}\right)\right)}_{D_{n-1}}\right].
\end{aligned}$$

We first show that C_n is a constant that equals the desired quantity. Notice that for a fixed initial state $(s,t)\in\mathcal{S}\times\mathcal{T}$ and a fixed sequence $\mathbf{x}^{n-1}\in\mathcal{X}^{n-1}$ we have $\mathsf{T}_{\mathsf{c}}((s,t),\mathbf{x}^{n-1}) = \left(\mathsf{T}_1(s,\mathbf{x}^{n-1}),\mathsf{T}_2(t,\mathbf{x}^{n-1})\right)$ and hence

$$\begin{aligned}
C_n &= \sum_{s',t'} D_{\mathrm{KL}}\left(\left(\widetilde{P}_{g_1}\right)_{s',\cdot}\,\middle\|\,\left(\widetilde{P}_{g_2}\right)_{t',\cdot}\right)\cdot p_{12}\left\{\mathsf{T}_1\left(s,\mathbf{x}^n\right)=s',\mathsf{T}_2\left(t,\mathbf{x}^n\right)=t'\right\}\\
&= \sum_{s',t'} D_{\mathrm{KL}}\left(\left(\widetilde{P}_{g_1}\right)_{s',\cdot}\,\middle\|\,\left(\widetilde{P}_{g_2}\right)_{t',\cdot}\right)\cdot \mathbf{p}_{12_{(s',t')}}.
\end{aligned}$$

We next show that D_n converges in probability and in particular $D_n - D_{n-1}\to 0$. For a fixed initial state (s,t) and a sequence $\mathbf{x}^n$, consider the sequence of states $s,\mathsf{T}_1\left(s,\mathbf{x}^1\right),\mathsf{T}_1\left(s,\mathbf{x}^2\right),\ldots,\mathsf{T}_1\left(s,\mathbf{x}^n\right)$, and let $n_{s',x} = n_{s',x}(s)$ denote the number of indices i such that $\mathsf{T}_1(s,\mathbf{x}^{i-1}) = s'$ and $x_i = x$. We have for all $t\in\mathcal{T}_s$

$$p_{12}\left(\mathbf{x}^n|s,t\right) = \prod_{i=1}^{n}\mathsf{P}_1\left(\mathsf{T}_1\left(s,\mathbf{x}^{i-1}\right),x_i\right) = \prod_{s',x}\mathsf{P}_1\left(s',x\right)^{n_{s',x}} = 2^{\sum_{s',x} n_{s',x}\log\mathsf{P}_1\left(s',x\right)} = 2^{n\sum_{s',x}\frac{n_{s',x}}{n}\log\mathsf{P}_1\left(s',x\right)}.$$

Since the associated stochastic process on states is stationary and ergodic, we have $\frac{n_{s',x}}{n}\to\left(\mathbf{p}_{g_1}\right)_{s'}\mathsf{P}_1(s',x)$ in probability as $n\to\infty$, and hence $p_{12}\left(\mathbf{x}^n|s,t\right)\to 2^{-nH(g_1)}$ in probability and *independent* of the initial state s. This implies $\left(p_{12}\left(s,t\,|\mathbf{x}^n\right)\right)_{(s,t)}$ and $\left(p_{21}\left(s,t\,|\mathbf{x}^n\right)\right)_{(s,t)}$ converge in probability to the stationary distribution $\mathbf{p}_{12}$ and $\mathbf{p}_{21}$, respectively, which shows that D_n converges and hence the theorem follows. ∎