# Active Adversaries from an Information-Theoretic Perspective: Data Modification Attacks

Mahtab Mirmohseni
Department of Electrical Engineering,
Sharif University of Technology, Tehran, Iran
Email: mirmohseni@sharif.edu

Panos Papadimitratos
Networked Systems Security Group,
KTH Royal Institute of Technology, Stockholm, Sweden
Email: papadim@kth.se

***Abstract*—We investigate the problem of reliable communication in the presence of active adversaries that can tamper with the transmitted data. We consider a legitimate transmitter-receiver pair connected over multiple communication paths (routes). We propose two new models of adversary, a "memoryless" and a "foreseer" adversary. For both models, the adversaries are placing themselves arbitrarily on the routes, keeping their placement fixed throughout the transmission block. This placement may or may not be known to the transmitter. The adversaries can choose their best modification strategy to increase the error at the legitimate receiver, subject to a maximum *distortion* constraint. We investigate the communication rates that can be achieved in the presence of the two types of adversaries and the channel (benign) stochastic behavior. For memoryless adversaries, the capacity is derived. Our method is to use the typical set of the anticipated received signal for all possible adversarial strategies (including their best one) in a compound channel that also captures adversarial placement. For the foreseer adversaries, which have enhanced observation capabilities compared to the memoryless ones, we propose a new coding scheme to *guarantee* resilience, i.e., recovery of the codeword independently of the adversarial (best) choice. We derive an achievable rate and we propose an upper bound on the capacity. We evaluate our general results for specific cases (e.g., binary symbol replacement or erasing attacks), to gain insights.**

## I. Introduction

Operation in adverse networks requires secure and reliable communication: data modifications should not be merely detected but data should be delivered (decoded correctly) at their destination. Cryptographic primitives can ensure detection but not correction and thus data delivery. Consider a general network connecting a Transmitter (Tx) - Receiver (Rx) pair over multiple disjoint communication paths (e.g., multiple frequency bands or antennas in wireless networks, or multiple routes in multi-hop networks); adversaries can be present in a number of those paths. The challenge is how to leverage the available alternative paths in order to achieve reliable communication in the presence of the adversary. What is the best one can do against a powerful adversary? More generally, what is the best communication rate one can achieve in the face of malicious faults (adversarial modifications) *and* benign faults (due to the communication channel stochastic behavior)?

Facets of this problem were addressed in the literature. One approach leverages cryptographic primitives to detect modifications and attempt retransmissions over alternative communication paths (while introducing redundancy to tolerate faults) [1]. This, however, does not address the fundamental limits of the system performance. Without cryptographic assumptions, the minimum needed connectivity is derived for resilient communication for a Tx-Rx pair over $n$ disjoint paths, termed *wires*, and disrupted by active adversaries that compromise a subset of these wires (the scenario is termed the *Dolev model*) [2]. The analysis in [2] does not consider communication rates and thus does not even attempt to achieve the best performance; it does not model channel noise and does not consider adversarial limitations or fine-grained actions.

In contrast, confidentiality received significant attention, notably after Wyner's seminal paper [3], with the majority of works concerned with passive eavesdroppers [4, Chapter 22]. Less attention, in an information-theoretic sense, was paid to *active* adversaries that modify the channel input of the legitimate transmitter. An early characteristic model is the Arbitrarily Varying Channel (AVC) [5], which assumes worst-case interference: the adversary controls the channel state to maximize the error probability at the receiver. Depending on what the adversary knows and the common randomness of the legitimate nodes, the capacity can differ considerably [6], [7]. However, it is not easy to translate erasing and replacement attacks to the AVC worst-case interference notations. In particular, AVC cannot capture data modification attacks or network structure, e.g., as the Dolev model does [2]. Given that confidentiality (passive adversaries) is broadly researched in the information-theoretic sense (also in [2]), *the challenge is how to achieve (secure and) reliable communication in the presence of active adversaries, in addition to channel noise, and derive fundamental limits of the capacity?*

In this paper, we address this challenge. We propose a novel information-theoretic setup that captures network structure, fine-grained and strong, yet realistic active adversarial behavior, along with channel stochastic behavior. We consider a Tx-Rx pair communicating across a number of disjoint paths (routes). The adversaries compromise a fixed number of these routes, thus they get access to the respective (noiseless) transmitted signals. The adversaries can choose their best strategy (knowing the transmitted signal) to modify and increase the error at the Rx. However, their mapping is subject to a maximum distortion constraint, i.e., a *distortion limit*. This limit, given a distortion measure (depending on the specific attack), determines the distance between the transmitted codeword and its modified version; e.g., for an erasing attack on binary transmissions, the percentage of bits the adversary can erase. The adversaries' placement (on the routes) is arbitrary but fixed throughout one transmission block; moreover, it may be known to the Tx. The adversaries' observations (of the transmitted signal) can be either instantaneous or cover the entire codeword. We propose accordingly two adversary types: *memoryless* and *foreseer*. Our goal is to find the reliable communication rate a Tx-Rx can achieve in the presence of either

978-1-4799-5186-4/14/$31.00 ©2014 IEEE

of these two types of adversary. Our average distortion limit and the consideration of channel stochastic behavior (noise) on top of adversarial faults lead to a generalized model compared to the Dolev one for active adversaries. The channel noise we introduce in our model, which allows us to take into account benign faults and noisy observations, is not taken into account in the Dolev model. The distortion limit allows practical assumptions, e.g., adversaries with noisy observations, with tactics to remain undetected, limited resources or time or attempts to mount an attack, or even cryptographic integrity protection for parts of the messages (e.g., immutable fields). By setting the noise to zero and the distortion limit to its maximum, we reduce our model to the Dovel model.

We derive the capacity for the memoryless adversaries. For the achievability part, we use a compound channel to model the adversaries' placement. For each compromised route, we consider the typical set of the anticipated received signals in all possible adversarial scenarios (including the one for the best adversarial strategy), subject to the distortion limit. Then, for the foreseer adversaries, we propose a coding scheme using two techniques: (i) the Hamming approach [8], to cope with worst-case errors inflicted by adversaries with access to the entire codeword, and (ii) a random coding argument, to recover from the channel stochastic noise. For the former, we use the Varshamov construction [9], to guarantee the required minimum distance needed to mitigate adversary-inflicted errors. Moreover, we obtain an upper bound to the capacity, taking an approach similar to that for the Hamming bound (i.e., limiting the volume of the Hamming balls). Finally, we gain insights through three special cases: replacement and erasing attacks on binary transmission and Gaussian jamming. We determine the proper distortion measures and channel distributions to model attacks that correspond to realistic situations, e.g., bit or packet replacement and dropping (selective forwarding), and evaluate our derived rates for those. For these cases, we consider explicitly the best adversarial strategy: we show the adversaries can achieve the lower bounds on the capacity we derived without specific assumptions on the adversary strategy. Our results for these special cases reveal that (i) knowing the adversaries' placement at Tx is not useful in terms of the achievable reliable rate, (ii) memory helps the adversaries significantly, and (iii) differentiates the foreseer effect from channel noise; while the memoryless effect is equivalent to channel noise.

## II. Channel Model

**Notation**: Upper-case letters (e.g., $X$) denote Random Variables (RVs) and lower-case letters (e.g., $x$) their realizations. $X_i^j$ indicates a sequence of RVs $(X_i, X_{i+1}, ..., X_j)$; we use $X^j$ instead of $X_1^j$ for brevity. The probability mass function (p.m.f) of a RV $X$ with alphabet set $\mathcal{X}$ is denoted by $p_X(x)$; occasionally subscript $X$ is omitted. The set of all possible distributions on $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$. $\pi(x, y|x^n, y^n) \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ shows the joint type (i.e., empirical p.m.f) of two sequences of length $n$, which can be extended to several $n$-length sequences. $\mathcal{P}_n(\mathcal{X}) \subset \mathcal{P}(\mathcal{X})$ consists all possible types of sequences $x^n \in \mathcal{X}^n$. For $q \in \mathcal{P}^n(\mathcal{X})$, the type class is defined as $\mathcal{T}^n(q) = \{x^n, p_X(x) = q\}$. $A_\epsilon^n(X, Y)$ is the set of $\epsilon$-strongly, jointly typical sequences of length $n$. $\mathcal{N}(0, \sigma^2)$ denotes a zero-mean Gaussian distribution with variance $\sigma^2$. $\mathcal{B}(\alpha)$ is a Bernoulli distribution with parameter $\alpha \in [0, 1]$. $\mathbb{F}_q$ is a finite field with $q$ elements. We define $[x]^+ = \max\{x, 0\}$. Unless specified, logarithms are in base 2. Throughout the paper, $i$ and $j$ indices are used for time and route number, respectively. $H_q : [0, 1] \to \mathbb{R}$ is the Hilbert $q$-ary entropy function $H_q(x) = x\log_q(q-1) - x\log_q x - (1-x)\log_q(1-x)$. Bold letters are used to show the column vectors of length $n_r$, e.g., $\mathbf{x}^n = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{n_r,1} & \cdots & x_{n_r,n} \end{bmatrix}$ and $\mathbf{x}^n(j)$ shows its $j$th row.

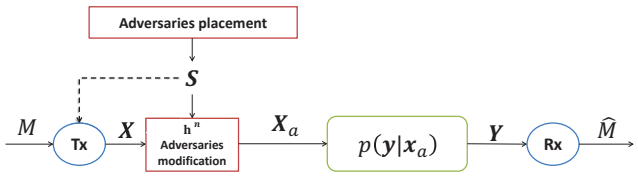


Fig. 1. Multi-route Point-to-Point channel with Modifying Adversaries (PP-MA).

**Channel model**: Consider a single unicast scenario: Tx sends a message $M$ to Rx, with $n_r$ available disjoint routes. $n_a$ out of the $n_r$ routes are attacked by the adversaries, with their placement being arbitrary but fixed throughout one transmission block. The placement can be chosen by the adversaries to maximize the error at Rx; but, it may be known to the Tx. One can implicitly assume there are $n_a$ adversaries: more than one adversary in a route can be modeled as a stronger adversary (i.e., with a higher distortion limit). We model this scenario with a (compound) state-dependent *multi-route Point-to-Point channel with Modifying Adversaries (PP-MA)* illustrated in Fig. 1: its transition probability is not entirely specified unless the *Channel State Information (CSI)* (i.e., adversaries' placement information) is known [10]. Consider finite alphabets $\mathcal{X}, \mathcal{X}_a, \mathcal{Y}$. The channel inputs at the Tx and the adversaries are defined by $\mathbf{X} \in \mathcal{X}^{n_r}$ and $\mathbf{X_a} \in \mathcal{X}_a^{n_r}$ respectively. $\mathbf{Y} \in \mathcal{Y}^{n_r}$ is the output of the channel at Rx. The $j$-th element of state vector $\mathbf{S} \in \{0, 1\}^{n_r}$, i.e., $\mathbf{S}(j)$, determines the presence of an adversary in $j$-th route. The received signal at Rx only depends on the adversary input, if present. Each adversary channel input must be relatively close to the Tx input in that route (subject to a distortion limit), according to some distortion metric. Hence, we define the $D$ class of adversaries for $j \in \{1, \ldots, n_r\}$ by the set of all probability distributions:

$$\mathcal{P}_j^a(D) = \{p_j(x_a^n|x^n) : E_{p_j^n}[d(X_a^n, X^n)] \le D\} \qquad (1)$$

where $d$ is a distortion measure defined by the mapping $d : \mathcal{X} \times \mathcal{X}_a \mapsto [0, \infty)$ and the average distortion for two sequences is $d(x_a^n, x^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_{a,i}, x_i)$. We assume the $\mathcal{X}_a \mapsto \mathcal{Y}$ channel is memoryless, thus the transition probability can be expressed by the conditional p.m.f on $\mathcal{Y} \times \mathcal{X}_a$ as:

$$\begin{aligned} p(\mathbf{y}^n, \mathbf{x}_a^n|\mathbf{x}^n, \mathbf{s}^n) &= p(\mathbf{y}^n|\mathbf{x}_a^n)p(\mathbf{x}_a^n|\mathbf{x}^n, \mathbf{s}^n) \qquad (2)\\ &= \prod_{j=1}^{n_r} p_j(\mathbf{y}^n(j)|\mathbf{x}_a^n(j))p_j(\mathbf{x}_a^n(j)|\mathbf{x}^n(j), \mathbf{s}^n(j)) \\ &= \prod_{j=1}^{n_r} p_j(\mathbf{x}_a^n(j)|\mathbf{x}^n(j), \mathbf{s}^n(j)) \prod_{i=1}^{n} p_j(\mathbf{y}_i(j)|\mathbf{x}_{a,i}(j)) \end{aligned}$$

The state vector, assumed fixed in one transmission block, models the channel statistics transmission block: $\mathbf{s}_i(j) = \mathbf{s}(j)$ for $i \in \{1, \ldots, n\}$ with at most $n_a \le n_r$ adversaries, i.e., $w_H(\mathbf{s}) \le n_a$. Hence, for $j \in \{1, \ldots, n_r\}$:

$$\begin{aligned} p_{j\,X_a|X,S}(\mathbf{x}_a^n(j)|\mathbf{x}^n(j), \mathbf{s}^n(j)) &= p_{j\,X_{a,s}|X}(\mathbf{x}_a^n(j)|\mathbf{x}^n(j)) \\ &= q_{j,s}(\mathbf{x}_a^n(j)|\mathbf{x}^n(j)) \end{aligned}$$

where

$$q_{j,s}(x_a^n|x^n) \in \mathcal{P}_j^a(D_{j,s} = s \cdot D_j) \quad (3)$$

which is due to the $D_j$ distortion limit at each adversary. In $n$ channel uses, Tx sends $M$ to Rx using the following code:

*Definition 1:* A $(2^{nR}, n, P_e^{(n)})$ code for the multi-route PP-MA consists of: (i) A message set, $\mathcal{M} = [1 : 2^{nR}]$, with message $M$ uniformly distributed over $\mathcal{M}$. (ii) An encoding function, $f^n$, at Tx, which maps $M$ to a codeword $\mathbf{x}^n \in \mathcal{X}^{n_r \times n}$. (iii) A set of adversaries' mapping, $\mathbf{h}^n$, with $h^n(j) : \mathcal{X}^n \times \{0,1\} \mapsto \mathcal{X}_a^n$ for $j \in \{1, \dots, n_r\}$ satisfying (3). (iv) A decoding function at Rx, $g : \mathcal{Y}^{n_r \times n} \mapsto \mathcal{M}$. (v) The probability of error for this code, defined as: $P_e^{(n)} = \frac{1}{2^{nR}} \sum_{m \in \mathcal{M}} Pr(g(\mathbf{y}^n) \neq m | m \text{ sent})$. In case the CSI is available at the Tx, we have: $f^n : \mathcal{M} \times \{0,1\}^{n_r} \mapsto \mathcal{X}^{n_r \times n}$. All codewords are revealed to all nodes (including adversaries). However, the adversaries' mapping is not known to the legitimate Tx and Rx.

*Definition 2:* A rate $R$ is achievable if there exists a sequence of $(2^{nR}, n, P_e^{(n)})$ codes such that for $\forall \mathbf{s} \in \{0,1\}^{n_r} : w_H(\mathbf{s}) \leq n_a$ and $\forall \mathbf{h}^n$ we have $P_e^{(n)} \to 0$ as $n \to \infty$. The capacity, $\mathcal{C}$, is the supremum of all achievable rates $R$.

**Memoryless active adversary:** Mapping at each adversary is:

$$p_j(\mathbf{x}_a^n(j)|\mathbf{x}^n(j), \mathbf{s}^n(j)) = \prod_{i=1}^{n} p_j(\mathbf{x}_{a,i}(j)|\mathbf{x}_i(j), \mathbf{s}_i(j)) \quad (4)$$

i.e., the adversary uses the same probability distribution to modify the transmitted symbols in each channel use. For each route $j$, the distribution in (4) is independent and identically distributed (i.i.d) and fixed over time; but, the distributions can differ across routes.

**Foreseer active adversary:** It observes the transmitted codeword over the entire block (i.e., $\mathbf{x}^n(j)$) upon which it bases its strategy. That is, while satisfying (3), the adversary can choose the position and value of the symbols in the codeword to be modified. In this case, we concentrate on two types of attacks:

*Replacement attacks*: $\mathcal{X} = \mathcal{X}_a = \mathcal{Y}$ with hamming distortion measure: $d(x, \hat{x}) = 1$ if $x \neq \hat{x}$ and $d(x, \hat{x}) = 0$ if $x = \hat{x}$.

*Erasing (dropping) attacks* (also known as selective forwarding): $\mathcal{X}_a, \mathcal{Y} = \{\mathcal{X}, e\}$ where for all $x, x' \in \mathcal{X}, x \neq x'$, $d(x, x) = 0$, $d(x, x') = \infty$ and $d(e, x) = d(x, e) = 1$. With this definition, we limit the adversaries only to erase the data and they cannot replace data as long as their distortion limits are finite, i.e., $D_j < \infty$.

These two types cover all possible modification attacks. It is reasonable to assume that anything outside the alphabet is rejected by Rx; thus, this can be modeled as an erased symbol. Therefore, the adversary does not gain anything by modifying to a non-existent symbol.

## III. Main Results

For our multi-route PP-MA, for both memoryless and foreseer adversaries, we consider either no CSI or CSI at Tx. The adversaries are assumed to have perfect CSI.

### A. Memoryless active adversaries

We state the capacity for the channel in (4), first assuming no CSI available at the Tx and Rx.

*Theorem 1:* The capacity of multi-route PP-MA satisfying (4), with no CSI available at either the Tx or the Rx is:

$$\mathcal{C}_{i.i.d}^{\text{nC}} = \sup_{p(\mathbf{x})} \min_{\substack{\mathbf{s} \in \{0,1\}^{n_r} \\ w_H(\mathbf{s}) \leq n_a}} \inf \sum_{j=1}^{n_r} I(\mathbf{X}(j); \mathbf{Y}_s(j)) \quad (5)$$

where infimum is taken over $\prod_{j=1}^{n_r} p_j(\mathbf{x}_a(j)|\mathbf{x}(j), \mathbf{s}(j)); \forall j \in \{1, \dots, n_r\} : E_{p_j}[d(\mathbf{X}_a(j), \mathbf{X}(j))] \leq D_{j,s} = s \cdot D_j$. $\forall \mathbf{s} \in \{0,1\}^{n_r}$, we have $\mathbf{Y}_s \in \mathcal{Y}^{n_r}$ and $p(\mathbf{y}^n|\mathbf{x}^n, \mathbf{x}_a^n, \mathbf{s}^n) = \prod_{j=1}^{n_r} p_{j\,Y|X_a,X,S}(\mathbf{y}^n(j)|\mathbf{x}^n(j), \mathbf{x}_a^n(j), \mathbf{s}^n(j)) = \prod_{j=1}^{n_r} p_{j\,Y_s|X_a,X}(\mathbf{y}^n(j)|\mathbf{x}^n(j), \mathbf{x}_a^n(j))$. Hence, the mutual information term is evaluated with respect to the joint p.m.f (2).

*Proof outline:* For the achievablility part, we use a random coding argument in a compound channel (to model the adversaries' placement). To take into account all possible i.i.d adversaries' strategies, we consider all possible joint types of $(\mathbf{x}^n(j), \mathbf{y}^n(j))$ for the $j$-th route, subject to the distortion limit on $\mathbf{x}_a^n(j)$. The converse follows from Fano's inequality, by noting that for *every* adversaries' placement and mapping ($\mathbf{s}$ and $\mathbf{h}^n$) we must have $P_e^{(n)} \overset{n \to \infty}{\longrightarrow} 0$. Detailed proof in [11]. ∎

*Remark 1:* On the $j$-th route, the conditional distribution of the adversary's channel input, i.e., $p_j(\mathbf{x}_a(j) \ |\mathbf{x}(j), \mathbf{s}(j))$, can model all possible memoryless active attacks (e.g., replacement or dropping). To specify a certain attack, it is enough to properly define the input alphabets, $\mathcal{X}_a, \mathcal{X}$, and the distortion measure $d(.,.)$. Thus, the $\inf$ is calculated over a feasible set of $\mathcal{X} \times \mathcal{X}_a$ distributions ($p_j$), where the feasibility constraint is determined by $E_{p_j}[d(\mathbf{X}_a(j), \mathbf{X}(j))] \leq D_{j,s} = s \cdot D_j$.

Next, we obtain the capacity when CSI is available at Tx (proof in [11]).

*Theorem 2:* The capacity of multi-route PP-MA satisfying (4), with CSI available at Tx is:

$$\mathcal{C}_{i.i.d}^{\text{TC}} = \min_{\substack{\mathbf{s} \in \{0,1\}^{n_r} \\ w_H(\mathbf{s}) \leq n_a}} \sup_{p(\mathbf{x})} \inf \sum_{j=1}^{n_r} I(\mathbf{X}(j); \mathbf{Y}_s(j)) \quad (6)$$

where the infimum is taken over $\prod_{j=1}^{n_r} p_j(\mathbf{x}_a(j)|\mathbf{x}(j), \mathbf{s}(j)); \forall j \in \{1, \dots, n_r\} : E_{p_j}[d(\mathbf{X}_a(j), \mathbf{X}(j))] \leq D_{j,s} = s \cdot D_j$, the notation $Y_s$ is defined in Theorem 1 and the mutual information term is evaluated with respect to the joint p.m.f (2).

### B. Foreseer active adversaries

Now, we derive lower and upper bounds on the capacity for all possible foreseer adversaries strategies. The bounds are based on the possible minimum distances the legitimate user codewords can tolerate under each attack.

*Theorem 3:* A lower bound to the capacity of the multi-route PP-MA with foreseer adversaries (no CSI available at Tx or Rx) is:

$$\mathcal{R}_l^{\text{nC}} = \sup \min \inf_{\mathbf{h}^n} \sum_{j=1}^{n_r} [H(\mathbf{V}) - H(\mathbf{X}_a(j)|\mathbf{Y}(j)) - \frac{H_{|\mathcal{X}|}(d_j)}{\log_{|\mathcal{X}|} 2}]^+$$

where the supremum and the minimum are taken over $p(\mathbf{x})p(\mathbf{v}|\mathbf{x}); \forall j \in \{1, \dots, n_r\} : E_{p_j}[d(\mathbf{V}(j), \mathbf{X}(j))] \leq d_j$ and $\mathbf{s} \in \{0,1\}^{n_r} : w_H(\mathbf{s}) \leq n_a$, respectively; $d_j = f(D_{j,s} = \mathbf{s}(j) \cdot D_j)$ is determined based on the attack type and the distortion measure (i.e., $d_j = \mathbf{s}(j) \cdot 2D_j$ for replacement attacks and $d_j = \mathbf{s}(j) \cdot D_j$ for erasing attacks); the second entropy, $H(\mathbf{X}_a(j)|\mathbf{Y}(j))$, is evaluated with respect to the memoryless channel: $p_j(\mathbf{y}^n(j)|\mathbf{x}_a^n(j)) = \prod_{i=1}^{n} p_j(\mathbf{y}_i(j)|\mathbf{x}_{a,i}(j))$.

*Proof outline:* We apply a random coding technique on top of a random linear code (Varshamov construction [9]), by introducing proper auxiliary codewords. Random coding is used

to combat the stochastic behavior of the $\mathcal{X}_a \mapsto \mathcal{Y}$ channel. Varshamov construction guarantees recovery from the worst-case errors, by making the minimum distance of the code greater than the number of errors. First, we generate auxiliary codewords, $\mathbf{u}$; then, we apply a random linear coding $n_r$ times to these codewords, to generate the transmitted codewords, $\mathbf{x}^n$. To decode from the $j$-th route: if Rx can decode the adversary's channel input $\mathbf{x}_a^n(j)$, the transmitted codeword is the only $\mathbf{x}^n(j)$ in a Hamming ball with radius $d_j$. To apply this scheme, we choose $\mathbf{v}^n(j)$ as the possible $\mathbf{x}_a^n(j)$ and try to decode it after receiving $\mathbf{y}^n(j)$, by decreasing its rate to satisfy the stochastic limitation imposed by the $\mathcal{X}_a \mapsto \mathcal{Y}$ channel. Proof details in [11]. ∎

*Theorem 4:* A lower bound on the capacity of the multi-route PP-MA with foreseer adversaries (CSI available at Tx) is:

$$\mathcal{R}_l^{\text{TC}} = \min \sup \inf_{\mathbf{h}^n} \sum_{j=1}^{n_r} [H(\mathbf{V}) - H(\mathbf{X}_a(j)|\mathbf{Y}(j)) - \frac{H_{|\mathcal{X}|}(d_j)}{\log_{|\mathcal{X}|} 2}]^+$$

where the minimum and the supremum are taken over $\mathbf{s} \in \{0,1\}^{n_r} : w_H(\mathbf{s}) \leq n_a$ and $p(\mathbf{x})p(\mathbf{v}|\mathbf{x}); \forall j \in \{1,\dots,n_r\} : E_{p_j}[d(\mathbf{V}(j), \mathbf{X}(j))] \leq d_j$, respectively; $d_j$ and $H_q(x)$ are defined in Theorem 3; the second entropy, $H(\mathbf{X}_a(j)|\mathbf{Y}(j))$, is evaluated with respect to $p_j(\mathbf{y}^n(j)|\mathbf{x}_a^n(j)) = \prod_{i=1}^{n} p_j(\mathbf{y}_i(j)|\mathbf{x}_{a,i}(j))$.

*Remark 2:* In both $\mathcal{R}_l^{\text{nC}}$ and $\mathcal{R}_l^{\text{TC}}$, the first term is independent of $\mathbf{s}$ and the second term is independent of $p(\mathbf{x})$. Therefore, we have $\mathcal{R}_l^{\text{nC}} = \mathcal{R}_l^{\text{TC}}$. That is, CSI does not help the achieving strategy for these rates.

*Theorem 5:* The following are upper bounds to the capacity of the multi-route PP-MA with foreseer adversaries:

$$\mathcal{R}_u^{\text{nC}} = \sup_{p(\mathbf{x})} \min \inf_{\mathbf{h}^n} \sum_{j=1}^{n_r} [I(\mathbf{X}_a(j); \mathbf{Y}(j)) - \frac{H_{|\mathcal{X}|}(\frac{d_j}{2})}{\log_{|\mathcal{X}|} 2}]^+ \quad (7)$$

$$\mathcal{R}_u^{\text{TC}} = \min \sup_{p(\mathbf{x})} \inf_{\mathbf{h}^n} \sum_{j=1}^{n_r} [I(\mathbf{X}_a(j); \mathbf{Y}(j)) - \frac{H_{|\mathcal{X}|}(\frac{d_j}{2})}{\log_{|\mathcal{X}|} 2}]^+ \quad (8)$$

where the minimum is taken over $\mathbf{s} \in \{0,1\}^{n_r} : w_H(\mathbf{s}) \leq n_a$ and $d_j$ and $H_q(x)$ are defined in Theorem 3.

*Proof outline:* We follow an approach similar to the one used to derive the Hamming bound, that is, we limit the volume of the coding balls. Proof in [11]. ∎

## IV. Examples

**Replacement attacks to binary transmission:** The channel inputs and output have binary alphabets (i.e., $\mathcal{X}, \mathcal{X}_a, \mathcal{Y} = \{0,1\}$) and $d$ is the Hamming distortion measure (defined in Section II). The stochastic channel from the adversary to the Rx is assumed to be a Binary Symmetric Channel (BSC). Thus, the channel output at Rx at time $i \in \{1,\dots,n\}$ is:

$$\mathbf{Y}_i(j) = \mathbf{X}_{a,i}(j) \oplus \mathbf{Z}_i(j) \quad (9)$$

where $\mathbf{Z}_i(j) \sim \mathcal{B}(N_j)$ for $j \in \{1,\dots,n_r\}$. First, consider memoryless active adversaries. We obtain the results of Theorems 1 and 2 as:

*Corollary 1:* The capacity of the multi-route PP-MA with binary alphabets satisfying (4), (9), for both no CSI and CSI at Tx, is:

$$\mathcal{C}_{i.i.d}^{\text{nC}} = \mathcal{C}_{i.i.d}^{\text{TC}} = n_r - \max_{\substack{\mathbf{s}\in\{0,1\}^{n_r} \\ w_H(\mathbf{s})\leq n_a}} \sum_{j=1}^{n_r} \max_{N_j' \leq \mathbf{s}(j)\cdot\tilde{D}_j} H(N_j * N_j')$$

where $\tilde{D}_j = \min\{D_j, 1-D_j\}$ and $\alpha * \beta = \alpha(1-\beta)+\beta(1-\alpha)$. If we assume identical route conditions, $D_j = D \leq \frac{1}{2}$ and $N_j = N$ for $j \in \{1,\dots,n_r\}$, the capacity is: $n_r - (n_r - n_a)H(N) - n_a H(N * D)$.

*Proof:* Let $P_j = Pr(\mathbf{X}(j) = 1)$ and without loss of generality assume $P_j \leq \frac{1}{2}$. To find the $\inf \sum_{j=1}^{n_r} I(\mathbf{X}(j); \mathbf{Y}_s(j))$ in (5), we first find a lower bound to it and we then show it is achievable by the adversaries.

$$I(\mathbf{X}(j); \mathbf{Y}_s(j)) = H(\mathbf{X}(j)) - H(\mathbf{X}(j)|\mathbf{Y}_s(j))$$
$$\geq H(P_j) - H(\mathbf{X}(j) \oplus \mathbf{X}_a(j) \oplus \mathbf{Z}(j)) \overset{(a)}{\geq} H(P_j) - H(N_j * N_j')$$

in (a) we define $N_j' \leq \mathbf{s}(j) \cdot D_j$ and use $Pr(\mathbf{X}(j) \neq \mathbf{X}_a(j)) \leq \mathbf{s}(j) \cdot D_j$. This lower bound is achievable by the $j$-th adversary if it chooses a joint distribution given by two backward BSCs, $\mathcal{Y} \to \mathcal{X}_a$ and $\mathcal{X}_a \to \mathcal{X}$, with cross-over probabilities $N_j$ and $N_j'$, respectively. This results in $Pr(\mathbf{X}_a(j) = 1) = \frac{P_j - N_j'}{1 - 2N_j'}$. Hence, we need $P_j \geq N_j'$ to hold. Therefore, (5) for this channel is:

$$\sup_{0\leq D_j \leq P_j \leq \frac{1}{2}} \min_{\substack{\mathbf{s}\in\{0,1\}^{n_r} \\ w_H(\mathbf{s})\leq n_a}} \sum_{j=1}^{n_r} [H(P_j) - \max_{N_j' \leq \mathbf{s}(j)\cdot D_j} H(N_j * N_j')]$$

The rest of the proof is straightforward. ∎

Now, consider foreseer active adversaries. We obtain the results of Theorems 3 and 4 as:

*Corollary 2:* The lower bound to the capacity of the multi-route PP-MA with foreseer adversaries, binary alphabets satisfying (9), for both no CSI and CSI at Tx is:

$$n_r - \sum_{j=1}^{n_r} H(N_j) - \max_{\substack{\mathbf{s}\in\{0,1\}^{n_r} \\ w_H(\mathbf{s})\leq n_a}} \sum_{j=1}^{n_r} H(\mathbf{s}(j) \cdot 2D_j) \quad (10)$$

For identical route conditions, $D_j = D \leq \frac{1}{2}$ and $N_j = N$ for $j \in \{1,\dots,n_r\}$, the rate is $n_r(1 - H(N)) - n_a H(2D)$.

*Proof:* Let $\mathbf{V}_i(j) = \mathbf{X}_i(j)$ and $P_j = Pr(\mathbf{X}(j) = 1)$. Recall that for all $\mathbf{h}^n$ (Definition 1, satisfying (3)), we have $Pr(\mathbf{X}(j) \neq \mathbf{X}_a(j)) \leq \mathbf{s}(j) \cdot D_j$. After some calculations, we can compute $H(\mathbf{V}(j)) = H(P_j)$ and

$$H(\mathbf{X}_a(j)|\mathbf{Y}(j)) \leq H(\mathbf{X}_a(j) \oplus \mathbf{Y}(j)) = H(\mathbf{Z}_i(j)) = H(N_j)$$

and obtain (3) as:

$$\mathcal{R}_l^{\text{nC}} \geq \sup_{0\leq P_j \leq \frac{1}{2}} \min_{\substack{\mathbf{s}\in\{0,1\}^{n_r} \\ w_H(\mathbf{s})\leq n_a}} \sum_{j=1}^{n_r} [H(P_j) - H(N_j) - H(\mathbf{s}(j) \cdot 2D_j)] \quad (11)$$

which will be maximized for $P_j = \frac{1}{2}$ independently of $\mathbf{s}(j)$, for $j \in \{1,\dots,n_r\}$. This results in (10). It is easy to see that computing $\mathcal{R}_l^{\text{TC}}$ in Theorem 4 results in the same rate. ∎

We adapt Theorem 5 for binary alphabets and the BSC of (10):

*Corollary 3:* The upper bound to the capacity of the multi-route PP-MA with foreseer adversaries, binary alphabets satisfying (9), for both no CSI and CSI at Tx, is:

$$n_r - \sum_{j=1}^{n_r} H(N_j) - \max_{\substack{\mathbf{s}\in\{0,1\}^{n_r} \\ w_H(\mathbf{s})\leq n_a}} \sum_{j=1}^{n_r} H(\mathbf{s}(j) \cdot D_j) \quad (12)$$

For identical route conditions, $D_j = D$ and $N_j = N$ for $j \in \{1,\dots,n_r\}$, the rate is $n_r(1 - H(N)) - n_a H(D)$.

*Proof:* We combine the methods of Corollaries 1 and 2. We can show that the sum of the first and the second terms in the right

side of (11) makes an upper bound on the first term of (7) and (8). To do this, it is enough to choose the proper joint distribution for the adversaries' input that achieves this bound. This distribution consists of two backward BSCs, $\mathcal{Y} \to \mathcal{X}_a$ and $\mathcal{X}_a \to \mathcal{X}$, with cross-over probabilities $N_j$ and $D_j$, respectively. The rest of the proof is similar to that of Corollary 2. ∎

**Erasing attacks on binary transmission:** To reduce the erasing attacks to binary alphabets, we set: $\mathcal{X} = \{0,1\}$, $\mathcal{X}_a, \mathcal{Y} = \{0,1,e\}$, $d(0,0) = d(1,1) = 0$, $d(0,1) = d(1,0) = \infty$, and $d(0,e) = d(1,e) = 1$. Across the $\mathcal{X}_a \mapsto \mathcal{Y}$ channel, additional erasing is introduced for the received signal at Rx (not distinguishable from the adversarial erasing at Rx). Thus, the channel output at Rx at time $i \in \{1,\dots,n\}$ is:

$$\mathbf{Y}_i(j) = \begin{cases} \text{BEC}(\mathbf{X}_{a,i}(j), N_j), & \text{if } \mathbf{X}_{a,i}(j) \neq e \\ \mathbf{X}_{a,i}(j), & \text{if } \mathbf{X}_{a,i}(j) = e \end{cases} \tag{13}$$

where $\text{BEC}(x,\beta)$ shows a Binary Erasure Channel (BEC) with input $x$ and probability of erasure $\beta$. Here, we state our results for both memoryless and foreseer adversaries. Proofs in [11].

*Corollary 4:* The capacity of the multi-route PP-MA with $\mathcal{X} = \{0,1\}$ and $\mathcal{X}_a, \mathcal{Y} = \{0,1,e\}$, satisfying (4) and (13), for both no CSI and CSI at Tx, is:

$$\mathcal{C}^{\text{nC}}_{i.i.d} = \mathcal{C}^{\text{TC}}_{i.i.d} = \min_{\substack{\mathbf{s}\in\{0,1\}^{n_r} \\ w_H(\mathbf{s})\le n_a}} \sum_{j=1}^{n_r} (1 - \mathbf{s}(j)\cdot D_j)(1-N_j).$$

For identical route conditions, $D_j = D$ and $N_j = N$ for $j \in \{1,\dots,n_r\}$, the capacity is $(1-N)(n_r - n_a D)$.

*Corollary 5:* The lower bound to the capacity of the multi-route PP-MA with foreseer adversaries, $\mathcal{X} = \{0,1\}$ and $\mathcal{X}_a, \mathcal{Y} = \{0,1,e\}$, satisfying (13), for both no CSI and CSI at Tx, is:

$$\min_{\substack{\mathbf{s}\in\{0,1\}^{n_r} \\ w_H(\mathbf{s})\le n_a}} \sum_{j=1}^{n_r} 1 - N_j(1-N'_j) - \bar{N}_j H(\frac{N'_j}{\bar{N}_j}) - H(\mathbf{s}(j)\cdot D_j)$$

where $\bar{N}_j = N_j(1-\mathbf{s}(j)\cdot D_j) + \mathbf{s}(j)\cdot D_j$. For identical route conditions, $D_j = D$ and $N_j = N$ for $j \in \{1,\dots,n_r\}$, the rate is $n_r(1-N) - n_a[(N(1-D)+D)H(\frac{D}{N(1-D)+D}) + H(D) - D]$.

*Corollary 6:* The upper bound to the capacity of the multi-route PP-MA with foreseer adversaries, $\mathcal{X} = \{0,1\}$ and $\mathcal{X}_a, \mathcal{Y} = \{0,1,e\}$, satisfying (13), for both no CSI and CSI at Tx, is:

$$\min_{\substack{\mathbf{s}\in\{0,1\}^{n_r} \\ w_H(\mathbf{s})\le n_a}} \sum_{j=1}^{n_r} [H((1-N'_j)(1-N_j)) \tag{14}$$
$$+(1-N'_j)(1-N_j-H(N_j)) - H(\mathbf{s}(j)\cdot\frac{D_j}{2})]$$

where $N'_j = \mathbf{s}(j)\cdot D_j$. For identical route conditions, $D_j = D$ and $N_j = N$ for $j \in \{1,\dots,n_r\}$, the rate is $n_a(H((1-N)(1-D)) + (1-D)(1-N-H(N)) - H(\frac{D}{2})) + (n_r - n_a)(1-N)$.

**Gaussian replacement attacks:** We assume Gaussian distributions for the channel inputs and output. The distortion measure now is the squared error distortion: $d(x,\hat{x}) = (x-\hat{x})^2$, and the channel model can be shown as:

$$\mathbf{Y}_i(j) = \mathbf{X}_{a,i}(j) + \mathbf{Z}_i(j) \tag{15}$$

where $\mathbf{Z}_i(j) \sim \mathcal{N}(0, N_j)$ for $j \in \{1,\dots,n_r\}$ are independent and i.i.d Gaussian noise components. We assume the average power constraint on input signal $\mathbf{X}(j)$ as $\frac{1}{n}\sum_{t=1}^{n} |\mathbf{x}_t(j)|^2 \le \mathbf{P}_j$.

TABLE I

| | | Replacement | Erasing |
|---|---|---|---|
| Memoryless | $\mathcal{C}$ | $1 - H(N*D)$ | $(1-N)(1-D)$ |
| Foreseer | $\mathcal{R}_l$ | $1 - H(N) - H(2D)$ | $1 - N(1-D) - H(D)$ $-(N(1-D)+D)H(\frac{D}{N(1-D)+D})$ |
| | $\mathcal{R}_u$ | $1 - H(N) - H(D)$ | $H((1-N)(1-D)) - H(\frac{D}{2})$ $+(1-D)(1-N-H(N))$ |

Hence, $\mathbf{X}(j) \sim \mathcal{N}(0, P_j)$ for $j \in \{1,\dots,n_r\}$. Here, we only consider the memoryless adversaries and obtain the results of Theorems 1 and 2 (proof in [11]).

*Corollary 7:* The capacity of the multi-route PP-MA with Gaussian distributions for channel inputs and output, satisfying (4) and (15), for both no CSI and CSI at Tx, is:

$$\mathcal{C}^{\text{nC}}_{i.i.d} = \mathcal{C}^{\text{TC}}_{i.i.d} = \max_{\substack{\mathbf{s}\in\{0,1\}^{n_r} \\ w_H(\mathbf{s})\le n_a}} \sum_{j=1}^{n_r} \theta(\frac{P_j - \mathbf{s}(j)\cdot D_j + N_j}{\mathbf{s}(j)\cdot D_j + N_j})$$

where $\theta(x) \doteq \frac{1}{2}\log(x)$. For identical route conditions, $D_j = D$ and $N_j = N$ with equal power constraints $P_j = P$ for $j \in \{1,\dots,n_r\}$, the capacity is: $n_r\theta(1+\frac{P}{N}) - n_a\theta(1+\frac{D(P+2N)}{N(P-D+N)})$.

**Comparison:** Along with identical route conditions, to simplify, let $n_r = n_a = 1$. Table I shows the results for the replacement and erasing attacks on binary transmission. Obviously, for zero distortion for the adversary ($D = 0$), we have BSC and BEC with parameter $N$. The rate reduction caused by a foreseer adversary is considerable. Consider only the adversary's effect by setting $N = 0$: the foreseer is twice more powerful than the memoryless one (in terms of the lower bound) for the replacement attack. For the erasing attack, the foreseer reduces (compared to the memoryless) the rate from a BEC rate (i.e., $1-D$) to a BSC rate (i.e., $1 - H(D)$). For Gaussian replacement attacks (under these simplified assumptions), the capacity is $\frac{1}{2}\log(1 + \frac{P-2D}{D+N})$; while, for Gaussian independent jamming with power $D$, we achieve $\frac{1}{2}\log(1 + \frac{P}{D+N})$. Thus, knowing the transmitted codeword (even in a memoryless case) worsens the situation compared to an independent jammer.

## REFERENCES


[1] P. Papadimitratos and Z. J. Haas, "Secure data communication in mobile Ad Hoc networks," *IEEE JSAC*, vol. 24, no. 2, pp. 343–356, Feb. 2006.

[2] D. Dolev, C. Dwork, O. Waarts, and M. Yung, "Perfectly secure message transmission," *J. ACM*, vol. 40, no. 1, pp. 17–47, Jan. 1993.

[3] A. D. Wyner, "The wire-tap channel," *Bell Systems Technical Journal*, vol. 54, no. 8, pp. 1355–1387, Jan. 1975.

[4] A. E. Gamal and Y.-H. Kim, *Network information theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[5] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacities of certain channel classes under random coding," *The Annals of Mathematical Statistics*, vol. 31, no. 3, pp. 558–567, 1960.

[6] H. Boche and R. F. Schaefer, "Capacity results, coordination resources, and super-activation in wiretap channels," in *Proc. IEEE Int. Symp. Info. Theory (ISIT)*, Istanbul, Turkey, Jul. 2013, pp. 1342–1346.

[7] A. D. Sarwate, "Coding against myopic adversaries," in *Proc. IEEE Information Theory Workshop (ITW)*, Dublin, Ireland, Aug. 2010.

[8] R. W. Hamming, "Error detecting and error correcting codes," *Bell System Technical Journal*, vol. 26, no. 2, pp. 147–160, 1950.

[9] R. Varshamov, "Estimate of the number of signals in error correcting codes," *Dokl. Akad. Nauk SSSR*, vol. 117, pp. 739–741, 1957.

[10] I. Csiszar and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. New York: Academic Press, 1982.

[11] M. Mirmohseni and P. Papadimitratos, "Active adversaries from an information-theoretic perspective: Data modification attacks," in arXiv:1404.6331 [cs.IT], April 2014.