Plausible Deniability over Broadcast Channels

Mayank Bakshi* Member, IEEE, and Vinod Prabhakaran[†] Member, IEEE

Abstract

In this paper, we introduce the notion of Plausible Deniability in an information theoretic framework. We consider a scenario where an entity that eavesdrops through a broadcast channel summons one of the parties in a communication protocol to reveal their message (or signal vector). It is desirable that the summoned party have enough freedom to produce a fake output that is likely plausible given the eavesdropper's observation. We examine three variants of this problem – Message Deniability, Transmitter Deniability, and Receiver Deniability. In the first setting, the message sender is summoned to produce the sent message. Similarly, in the second and third settings, the transmitter and the receiver are required to produce the transmitted codeword, and the received vector respectively. For each of these settings, we examine the maximum communication rate that allows a given minimum rate of plausible fake outputs. For the Message and Transmitter Deniability problems, we fully characterise the capacity region for general broadcast channels, while for the Receiver Deniability problem, we give an achievable rate region for physically degraded broadcast channels.

I. INTRODUCTION

The explosive growth in information technologies in recent years is not without its pitfalls. On one hand, advances in communications have enabled ground-breaking applications that have arguably been instrumental in improving the general quality of life. On the other hand, the naturally connected nature of these technologies also presents a wide variety of security and privacy concerns. To counter these, much recent attention has also focused on designing and analyzing algorithms and protocols that guarantee security or privacy. It is worth noting that the security requirement often varies greatly with the application. Indeed, the consequences of security failure as well as the nature of eavesdropping parties differ from application to a potentially malicious party. On the other hand, for an whistleblower posting sensitive information to an accomplice, any security failure has potentially life-altering consequences. The nature of the eavesdropper is also different in these situations. In the first example, an eavesdropper is typically a passive party that simply listens to an ongoing transmission, and it is desirable that the content of the communication be kept hidden from the eavesdropper. On the other hand, in the second example, the eavesdropper may often be an authority that has the power to *coerce* the whistleblower to reveal the transmitted message. In this case, it is important that the whistleblower is able to *deny* the fact that any sensitive communication has taken place by producing a fake message that appears plausible to the coercing party.

We argue that while much of the work in secure communication is well suited to the first scenario, *i.e.*, the ability to *hide* data, there is relatively little work that applies to the second scenario. For the first scenario, by now, there is are well developed theoretical results as well as practical algorithms both in the *cryptographic* [1] as well as *information theoretic* [2]–[4] settings. However, there is limited understanding of both fundamental limits and algorithms for the second setting. In this paper, we propose an information theoretic framework for *Plausibly Deniable* communication in the sense just described. In the following, we begin with an overview of some related notions of security and contrast these with our notion of Plausibly Deniable communication.

A. Related notions

1) Information theoretic secrecy: Usually secure protocols aim to hide data from an eavesdropper by taking advantage of some asymmetry between the legitimate receiver and the eavesdropper – the eavesdropper should be "less powerful" than the legitimate receiver. The framework of information theoretic secrecy relies on the eavesdropper having "less information" than the intended receiver and provides guarantees that hold irrespective of the eavesdropper's computational ability. For example, in the wiretap channel setting [2], [3] (See Figure 4a) the eavesdropper may observe Alice's transmission through a noisier channel than Bob does. On similar lines, in the secure network coding setting [5], the eavesdropper may observe a smaller subset of the transmission than legitimate nodes. In each of these settings, the information-theoretic approach allows characterizing the "capacity", which is defined as the maximum code rate such that (a) the intended receiver can decode the secret message m reliably given her received vector y, *i.e.*, $P(\hat{m}(y) \neq m) \approx 0$, and (b) the eavesdropper can gain very little statistical information about the secret message m given her observation z, *i.e.*, $P(m|z) \approx P(m)$. Note here that there is no restriction placed on

a Ramanujan fellowship from the Department of Science and Technology, Government of India and in part by Information Technology Research Academy (ITRA), Government of India under ITRA-Mobile grant ITRA/15(64)/Mobile/USEAADWN/01.

A preliminary version of this work was presented at the 2016 IEEE International Symposium on Information Theory, Barcelona, Spain.

^{*}Mayank Bakshi (*mayank@inc.cuhk.edu.hk*) is with the Institute of Network Coding, The Chinese University of Hong Kong. The work described in this paper was partially supported by a grant from University Grants Committee of the Hong Kong Special Administrative Region, China (Project No. AoE/E-02/08). [†]Vinod Prabhakaran (*vinodmp@tifr.res.in*) is with the Tata Institute of Fundamental Research, India. Vinod Prabhakaran's research was funded in part by



Fig. 1: Alice wishes to communicate a message *m* to Bob by sending a codeword **x** over a noiseless binary channel while an eavesdropper Judy observes **x** through a binary erasure channel with erasure probability p > 0. Note that, in order to avoid being detected as lying, the summoned party's output should appear plausible to Judy given her side information **z**. In particular, for the channel in this example, both Alice and Bob are forced to reveal their true codewords (*i.e.*, **x**) to Judy. This example also shows a contrast between the standard notion of secrecy and the plausible deniability requirement.

the computational power of the eavesdropper. As a result, schemes that guarantee information theoretic security are free of computational assumptions and as a result are guaranteed to be secure against any future developments in fast computing.

We argue that even though information theoretic secrecy is perfectly suited when the goal is to only hide the data against a passive eavesdropper, it does not guarantee any protection against eavesdroppers that have the ability to summon one of the communicating parties. The reason for this is as follows. At a high level, information theoretic secrecy is achieved by ensuring that the eavesdropper has a large enough list of candidate messages that appear roughly equiprobable. On the other hand, plausible deniability requires the summoned party to produce one such candidate message *without knowing* the eavesdropper's channel realisation. The following example illustrates this difference more concretely.

Example 1 (Secrecy does not guarantee plausible deniability). Consider the setting of Figure 1. Since the channel to Bob is noiseless, the secrecy capacity [3] is p. On the other hand, even if Alice and Bob operate a code equipped with an information-theoretic secrecy guarantee and Judy *demands* that Alice provide the transmitted codeword \mathbf{x} , Alice has no choice but to provide exactly what was transmitted (and hence, also reveal the message). If Alice chooses to provide a vector \mathbf{x}' different from \mathbf{x} , then Judy would be able to detect with a constant probability that Alice is lying since the transmitted symbol for any coordinate where \mathbf{x}' and \mathbf{x} differ would be received correctly by Judy with probability 1 - p.

2) Cryptographic security: In the cryptographic setting, the asymmetry between the legitimate receiver and the eavesdropepr usually manifests itself through complexity theoretic notions. For example, in a *public key cryptosystem*, the receiver holds a pair of carefully chosen keys (k_{public} , $k_{private}$). The public key k_{public} is known to all parties including the eavesdropper, while the private key $k_{private}$ is known only to the eavesdropper. This allows the sender to encrypt the message *m* to the ciphertext $\mathbf{x} = \text{ENC}(m, k_{public})$. The encryption algorithm is chosen such that the receiver can use his private key to decrypt the ciphertext to obtain the message as $m = \text{DEC}(\mathbf{x}, k_{public}, k_{private})$ in polynomial time. On the other hand, without knowing $k_{private}$, the eavesdropper cannot efficiently compute $\text{ENC}^{-1}(\mathbf{x}, k_{public})$ (under reasonable computational assumptions). However, even if the eavesdropper is unable to invert the ciphertext on their own, if they have the ability to summon the receiver to produce the private key, the receiver may have no choice but to respond truthfully by revealing the true private key, else the ciphertext and the public key may not be consistent with it.

3) Deniable Encryption: The notion of Deniable Encryption was first introduced by Canetti *et al.* in [6] recognizing the above problem of lack of plausible deniability in the cryptographic setting.¹ Here, the typical setting is as follows. Consider a public key setting as described in Section I-A2. Unlike the setting of Section I-A2 the eavedropper Judy who has bounded computational power both observes the ciphertext and can issue a summon to Bob coercing him to revealing the message. The framework of Deniable Encryption allows for encryption schemes such that upon receiving Judy's summon, Bob is able to produce a fake private key $k_{\text{public}}^{(\text{re})}$ which decrypts the ciphertext to a fake message $m^{(\text{re})}$ while appearing plausible to Judy. In other words, there is no polynomial time algorithm, using which Judy is able to determine whether Bob has responded with the true public key or a fake public key. Note that usual public key protocols such as RSA do not allow Bob to produce a fake key for every pair of (m, k_{public}) . This notion has received much attention in recent years. By now, there are fairly extensive theoretical and practical developments along this line (c.f. [8]-[10] and the references therein).

4) Covert Communication: In both the secrecy and the plausible deniability problems considered above, while the goal is to be able to hide the message that is being transmitted, the implicit assumption is that it is permissible for some form of communication to take place. However, in the setting of *covert communication* [11]–[14], even the fact that any communication is taking place is objectionable from the eavesdropper's point of view. For example, the communicating parties may be two prisoners in adjacent cells that wish to communicate without the warden knowing that they are doing so. In this setting, the goal is to ensure that from the warden's point of view, the output distribution induced by non-zero transmissions appear close

to that under zero transmission. The capacity for this problem is now well understood and follows the so called *square-root* law – in *n* channel uses, only $O(\sqrt{n})$ message bits can possibly be transmitted without being detected. Note that the notion of covertness only guarantees that the eavesdropper be unable to distinguish no transmission from a non-zero transmission; it does not necessarily prevent the eavesdropper from gaining any information about the potential message, if she assumes that something was transmitted.² Therefore, the covertness requirement only implies a weak form of plausible deniability – the transmitter can claim that no transmission took place when something was transmitted. However, it does not necessarily allow the communicating parties to claim the transmission of a message different from the true message.

B. Our work

Taking inspiration from the formulation of Deniable Encryption discussed in Section I-A3, we propose an information theoretic approach to plausible deniability. While the approach in Section I-A3 relies on cryptographic assumptions, *i.e.*, the assumption that the eavesdropper is computationally limited without access to the receiver's private key, we assume that the eavesdropper has potentially unlimited computational power, but the eavesdropper and the legitimate receiver have different channels statistics. In this setting, the sender can leverage this difference by careful encoding that allows the receiver to decode the message correctly while leaving enough room for confusion such that, if summoned, transmitter and the receiver are able produce fake messages or codewords that appear statistically indistinguishable from the true message or codeword to the eavesdropper given his channel observation.

1) Our setup: Our general setup is as follows. Alice, Bob, and Charlie are three participants in a potentially secretive communication setup. Charlie wishes to send a message $m \in \mathcal{M}$ to Bob through Alice. Alice and Bob are at two ends of a noisy channel and operate the physical layer with Alice being the transmitter and Bob being the receiver, while Charlie interacts directly with Alice and knows the message but does not partake in the physical layer transmission and reception. The nature of the message may either be an innocuous or a secretive one – this is known to Alice, Bob, and Charlie, but not to any eavesdroppers.

Judy is an eavesdropper who observes a noisy version of Alice's transmission. In this work, we assume that the statistics of Judy's observation are known to the above three parties, but the exact observation is unknown. We consider three settings for this problem. In the Transmitter Deniability problem, Judy may summon Alice and ask her to produce the transmitted codeword. Similarly, in the Receiver Deniability, and the Message Deniability problems, Judy may summon Bob, and Charlie, to produce the received vector, and the message, respectively. In each of these settings, depending on whether the communication is innocuous or secretive, the summoned party may either respond truthfully or use a *Faking Procedure* to produce a fake output that reveals as little information about the true message as possible while still maintaining plausibility with respect to Judy's observation.

We quantify the efficacy of a communication scheme in terms of its two properties – the *reliability* of the code and the *plausible deniability* of the faking procedure. The first property *i.e.*, the reliability is measured in a standard fashion in terms of the *message rate* and the *error probability* at the decoder. Plausible deniability is also measured in terms of two metrics – the *plausibility* and the *rate of deniability*. Roughly speaking, plausibility measures the closeness between two distributions – the joint distribution of the fake output with the eavesdropper's observation and that of the true message or signal vector with the eavesdropper's observation. We measure this distance in terms of the Summoned party's observations. This attempts to capture the amount of freedom the summoned party has while responding to the summoned party is forced to respond. Strictly speaking, the rate of deniability is a purely operational characteristic of the faking procedure and our formal definition of the rate of deniability does not appear to be related to equivocation. However, when the faking procedure satisfies the plausibility requirement, we establish an asymptotic equivalence between these two notions in Propositions 2 and 3. We also emphasise here that demanding a rate of deniability *D* is a stronger requirement than demanding an equivocation *D* in the usual information theoretic secrecy setting – this naturally extends similar observations in the cryptographic setting where, a plausibly deniable protocol trivially also satisfies the security requirement.

2) Organization of this paper: The rest of this paper is organised as follows. In Section II, we formally describe our notation and problem formulation and state the main results in Section III. In Sections IV and V, we give proof sketches for our theorems, and discuss some examples and key properties of our capacity regions. Finally, in Section VI, we provide concluding remarks.

²One can also demand both covertness and secrecy simultaneously. By operating at even lower rates (though still $O(\sqrt{n})$ bits per *n* channel uses), it is possible to be covert about the transmission status and secret about the message being potentially transmitted. [13], [15].

³Although, in this paper, we measure the plausibility in terms of K-L divergence, one is also well justified to instead use other measures of distance such as the variational distance. We argue that K-L divergence is a stronger measure for our problem as requiring that the K-L divergence be small also implies that the variational distance is small (by invoking Pinsker's inequality). Further, using K-L divergence instead of variational distance considerably simplifies our converse proofs. It is worth noting that the variational distance has a natural interpretation in terms of Hypothesis Testing – the variational distance between two probability measures P₁ and P₂ equals $1 - Pr(\text{test outputs P}_2|\text{true distribution is P}_1) - Pr(\text{test outputs P}_1|\text{true distribution is P}_2)$ for an optimal hypothesis test for distinguishing P₁ and P₂.

II. PROBLEM FORMULATION

A. Notation

Throughout this paper, we typically adopt the following notation. Upper case math and lower case symbols such as X and x denote random variables and their specific values respectively. Boldface symbols such as X and x denote random vectors and their specific values respectively, while calligraphic symbols such as \mathscr{X} denote sets. Probability distributions of generic random variables is typically written as P (e.g. P_X , $P_{Y|X}$), while probability distributions imposed by the specific codebook are typically written as Q (e.g. Q_X). All logarithms in this paper are assumed to base 2. For some random variables X and Y following distributions P_X and P_Y on alphabets \mathscr{X} and \mathscr{Y} respectively, we define the entropy, conditional entropy, and the mutual information respectively as $H(X) \triangleq \sum_{x \in \mathscr{X}: P_X(x) > 0} P_X(x) \log (1/P_X(x))$, $H(Y|X) \triangleq \sum_{x \in \mathscr{X}: P_{X,Y}(x,y) > 0} P_{X,Y}(x, y) \log (1/P_{Y|X}(y|x))$, and I(X; Y) = II(X) - II(X|Y). The Kullback-Leibler divergence between two probability measures P_1 and P_2 over a set \mathscr{X} is defined as $D(P_1||P_2) \triangleq \sum_{x \in \mathscr{X}: P_1(x) > 0} P_1(x) \log (P_1(x)/P_2(x))$. Throughout this paper, we employ strong typicality in our analysis, and define the strongly typical set for a random variable X as

$$\mathscr{A}_{\epsilon}^{(n)}(X) \triangleq \left\{ \mathbf{x} \in \mathscr{X}^n : \max_{x \in \mathscr{X}} \left| \frac{|\{i : x_i = x\}|}{n} - \mathsf{P}_X(x) \right| \le \frac{\epsilon}{|\mathscr{X}|} \right\}.$$

B. Channel model

Consider the problem settings shown in Figure 2. Alice, Bob, and Judy are connected through the following memoryless broadcast channel – at each discrete time instant, Alice's transmission $X \in \mathcal{X}$, Bob's reception $Y \in \mathcal{Y}$, and Judy's observation $Z \in \mathcal{Z}$ follow the conditional distribution $\mathsf{P}_{Y,Z|X}$ over finite alphabets $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Initially, only Charlie knows the message $m \in \mathcal{M}$ and passes it onto Alice to be transmitted to Bob over the broadcast channel. Charlie only knows the value of the message, but does not see the channel inputs or outputs. Throughout this paper, we assume that the message M is uniformly distributed over \mathcal{M} . There is no shared randomness, but Alice, Bob, and Charlie have private randomness $K_A \in \mathcal{K}$, $K_B \in \mathcal{K}$, and $K_C \in \mathcal{K}$ respectively. In addition, the code and the faking procedure (defined in the following) are known to all parties.

C. Codes and Faking Procedures

A code of block-length *n* is a pair of maps ENC : $\mathscr{M} \times \mathscr{K} \to \mathscr{X}^n$ and DEC : $\mathscr{Y}^n \to \mathscr{M}$. These maps are applied by Alice and Bob to generate the codeword $\mathbf{x} \triangleq x^n = \text{ENC}(m, k_A)$ and the reconstruction $\hat{m} = \text{DEC}(\mathbf{y})$ respectively. When there is no private randomness at Alice, we denote the codeword for message *m* by $\mathbf{x}(m)$. To simplify notation, we represent a code (ENC, DEC) through its codebook $\mathscr{C} \triangleq \{\text{ENC}(m, k_A) : m \in \mathscr{M}, k_A \in \mathscr{K}\}$. Note that \mathscr{C} is a multi-set with possible repetitions as we do not require that $\text{ENC}(\cdot)$ be an injective map.

Judy may summon Alice, Bob, or Charlie to provide a variable $\mathbf{w} \in \mathcal{W}$ that can be used to reconstruct the message using a map MSG : $\mathcal{W} \to \mathcal{M}$. Depending on whether or not the transmission is an innocuous, the summoned party may either reveal the true value of \mathbf{w} or use a (possibly stochastic) *faking procedure* FAKE : $\mathcal{W} \times \mathcal{K} \to \mathcal{W}$ to output a fake value $\mathbf{w}^{(F)} \in \mathcal{W}$. In this paper, we consider three settings that are specified by the choice of the variable \mathbf{w} . In particular, we consider the following special cases:

a) Message deniability: This setting is shown in Figure 2a. Charlie is the summoned party, $\mathbf{w} = m$, $\mathcal{W} = \mathcal{M}$, and $MsG(\mathbf{w}) = \mathbf{w}$.

b) Transmitter deniability: This setting is shown in Figure 2b. Here, Alice is the summoned party, $\mathbf{w} = \mathbf{x}$, $\mathcal{W} = \mathcal{X}^n$, and MSG(\mathbf{w}) is the most likely message given that $\mathbf{x} = \mathbf{w}$, *i.e.*, MSG(\mathbf{w}) $\triangleq \operatorname{argmax}_{m \in \mathcal{M}} Q_{M|\mathbf{X}}(m|\mathbf{w})$ if the maximum is attained at a unique value of *m*. If there are multiple values of *m* achieving the above maximum, then MSG(\mathbf{w}) selects one of them arbitrarily.

c) Receiver deniability: This setting is shown in Figure 2c. Bob is the summoned party, $\mathbf{w} = \mathbf{y}$, $\mathcal{W} = \mathcal{Y}^n$, and $MsG(\mathbf{w}) = DEC(\mathbf{w})$.

D. Reliability

We say that \mathscr{C} is (ϵ, R) -reliable if $\frac{1}{n} \log |\mathscr{M}| = R$, and there exists an encoder and decoder pair (ENC, DEC) such that the average error probability $\sum_{(m,\mathbf{y}): DEC(\mathbf{y}) \neq m} Q_{M,\mathbf{Y}}(m,\mathbf{y})$ is no larger than ϵ . Here, Q_M is the uniform distribution on \mathscr{M} and $Q_{\mathbf{Y},M}$ is the joint distribution of the message M and Bob's received vector \mathbf{Y} that induced by the specific code (ENC, DEC) and the channel transition probability $P_{YZ|X}$.

E. Plausible deniability

We first define our notion of plausible deniability for general random variables, and subsequently, specialise it to our setting. Let $\mathbf{W}^{(\mathrm{F})}$, \mathbf{W} , and \mathbf{Z} be random variables distributed according to a distribution $\mathbf{Q}_{\mathbf{W}^{(\mathrm{F})},\mathbf{W},\mathbf{Z}}$. Let $\mathbf{Q}_{\mathbf{Z},\mathbf{W}}$ and $\mathbf{Q}_{\mathbf{Z},\mathbf{W}^{(\mathrm{F})}}$ be marginals of the distribution $\mathbf{Q}_{\mathbf{Z},\mathbf{W},\mathbf{W}^{(\mathrm{F})}}$. We say that $\mathbf{W}^{(\mathrm{F})}$ is (δ, D) -plausibly deniable for \mathbf{W} given observation \mathbf{Z} if

(i) $\mathbb{D}(\mathsf{Q}_{\mathbf{Z},\mathbf{W}^{(F)}}||\mathsf{Q}_{\mathbf{Z},\mathbf{W}}) \leq \delta$, and



(a) Message deniability



(b) Transmitter Deniability



(c) Receiver Deniability

Fig. 2: The above figure shows the three different problem settings considered in this paper. These settings have the following commonalities: Charlie knows only the message *m* and may have access to an independently generated private random string k_C ; Alice knows the message *m*, an independently generated private random string k_A and the transmitted codeword **x**; Bob observes the channel output **y** and potentially has an independently generated private random string k_B , and is required to reconstruct *m*; Judy observes the channel output **z**. However, depending on the setting we consider, Judy summons Charlie, Alice, or Bob to produce *m*, **x**, or **y** respectively. The summoned party responds with a fake output FAKE(·) that has roughly the same distribution as the variable Judy demands to know. In each setting, the fake output is a function of the true value of variable demanded and the independent private randomness available to the summoned party.

(ii) $\frac{1}{n}\mathbb{H}(MSG(\mathbf{W}^{(F)})|\mathbf{W}) = D.$

In this paper, we are interested in settings where **W** is the random variable whose value is demanded by Judy through her summon, $\mathbf{W}^{(r)}$ is the random variable denoting the output of the faking procedure FAKE(·) employed by the summoned party, and **Z** is Judy's observation. The parameters δ and D respectively measure the *plausibility* and the *rate of deniability* of FAKE(·). We say that a faking procedure FAKE(·) is (δ, D) -plausibly deniable for **W** given observation **Z** is its output $\mathbf{W}^{(r)}$ is (δ, D) -plausibly deniable for **W** given observation **Z**.

Remark 1. Note that since we assume that the output of the faking procedure depends only value of variable W (that is known to the summoned party) and the summoned party's independently distributed private randomness, the random variables $W^{(F)}$, W, Z satisfy the Markov chain $W^{(F)} - W - Z$.

Remark 2. Note that the joint distribution $Q_{Z,W,W^{(F)}}$ depends on both the code (ENC, DEC) and the faking procedure, FAKE(·) and takes into account the (uniform) message distribution Q_M , the channel conditional probability $P_{YZ|X}$, and the distribution of independent private randomness variables K_A , K_B , and K_C .

F. Capacity regions

For each setting $\mathbf{w} \in \{m, \mathbf{x}, \mathbf{y}\}$, we say that a rate-deniability pair (R, D) is achievable if for any $\epsilon, \delta > 0$, for some $R' \ge R$ and $D' \ge D$, and for large enough *n*, there exists a blocklength-*n* code \mathscr{C} that is (ϵ, R') -reliable and a faking procedure FAKE(·) that is (δ, D') -plausibly deniable for W given Z. The capacity region $\mathscr{R}_{\mathbf{w}}$ is the closure of the set of all achievable rate-deniability pairs.

III. MAIN RESULTS

For the message deniability problem, we give a characterisation the capacity region \mathscr{R}_m for general broadcast channels in Theorem 1. The proof of this theorem is presented in Section IV.

Theorem 1 (Message Deniability). \mathscr{R}_m is the set of all (R, D) pairs such that

$$0 \le R \le \mathbb{I}(Y; V) + \mathbb{I}(U; Y|V) - \mathbb{I}(U; Z|V), \text{ and}$$

$$0 \le D \le \min \{R, \mathbb{I}(U; Y|V) - \mathbb{I}(U; Z|V)\}$$

for some random variables U and V which take values in sets \mathscr{U} and \mathscr{V} , respectively, with $|\mathscr{U}| \le (|\mathscr{X}| + 1)(|\mathscr{X}| + 2)$ and $|\mathscr{V}| \le |\mathscr{X}| + 2$, and satisfy the Markov chain V - U - X - (Y, Z).

Next, we characterise the capacity region \mathscr{R}_x for the transmitter deniability problem for general broadcast channels and given an achievable region for the receiver deniability problem for physically degraded broadcast channels. These results are stated in Theorems 2 and 3 below and are proved in Section V.

Theorem 2 (Transmitter Deniability). $\mathscr{R}_{\mathbf{x}}$ is the set of all (R, D) pairs such that

$$0 \le R \le \mathbb{I}(X; Y), \text{ and}$$
$$0 \le D \le \min \{R, \mathbb{I}(X; Y|U)\}$$

for some random variable U which takes values in a set \mathscr{U} , with $|\mathscr{U}| \leq |\mathscr{X}|$, and satisfies the Markov chains U - X - (Y, Z) and X - U - Z.

Theorem 3 (Achievability for Receiver Deniability). Let $P_{Y,Z|X}$ be a physically degraded broadcast channel, *i.e.*, $P_{Z|X}(z|x) = \sum_{y \in \mathscr{Y}} P_{Z|Y}(z|y) P_{Y|X}(y|x)$ for some distribution $P_{Z|Y}$. Then, \mathscr{R}_y includes all (R, D) pairs such that

$$0 \le R \le \mathbb{I}(X; Y), \text{ and}$$
$$0 \le D \le \min \{R, \mathbb{I}(X; Y|V)\}$$

for some random variable V which takes values in a finite set \mathscr{V} and satisfies the Markov chains V - Y - (X, Z) and Y - V - Z.

IV. MESSAGE DENIABILITY

In this section, we outline the proof of Theorem 1 and discuss connections of the message deniability problem with standard information theoretic secrecy problems. Our achievability argument relies on reducing our problem to the following variant of the information theoretic secrecy problem.



Fig. 3: Any code for the above secrecy problem can be operated as a code for the Message Deniability problem by treating s as the part of the message that the faking algorithm randomizes over and t as the part of the message that is unchanged by it.

A. Broadcast channel with confidential and leaked messages

Consider the setup shown in Figure 3. Alice observes sources $s \in \mathscr{S}$ and $t \in \mathscr{T}$ and wishes to transmit them reliably to Bob over *n* uses of the channel. Judy observes a noisy version of the transmission and knows the source *t* as side information. The goal for the transmission is to ensure that the leakage $I(S; \mathbb{Z}|T)$ is small. At first sight, the setting here is similar to the public message and confidential message setting of [3] in that secrecy is only required for the private message *s*. However, in contrast to [3], Judy is not interested in estimating *t* based on \mathbb{Z} , but is instead provided with *t* as side-information. This allows us to operate at potentially higher rates than [3]. We define the capacity region for this problem in the following.

Definition 1. The capacity region \mathscr{R}_s for broadcast channel with confidential and side-information messages is the set of (R_s, R_t) pairs such that, given $\epsilon, \delta > 0$, a large enough blocklength *n*, and sources *S* and *T* drawn independently and uniformly from \mathscr{S} and \mathscr{T} respectively, there exists a code \mathscr{C} , consisting of an encoder ENC : $\mathscr{S} \times \mathscr{T} \times \mathscr{K} \to \mathscr{X}^n$, a decoder DEC : $\mathscr{Y}^n \to \mathscr{S} \times \mathscr{T}$, and Alice's private randomness $K \in \mathscr{K}$, that satisfies the following properties:

1) $|\mathscr{S}| \geq 2^{nR_s}$ and $|\mathscr{T}| \geq 2^{nR_t}$.

- 2) $Q_{S,T,\mathbf{Y}}(\text{DEC}(\mathbf{Y}) \neq (S,T)) \leq \epsilon$.
- 3) $\mathbb{I}(S; T, \mathbb{Z}) < \delta$.

The following lemma provides an inner bound on \mathcal{R}_s .

Lemma 1. \mathcal{R}_s includes the set of all (R_s, R_t) pairs such that there exist random variables U and V satisfying V - U - X - (Y, Z),

$$R_s \le \mathbb{I}(U; Y|V) - \mathbb{I}(U; Z|V), \text{ and}$$
(1)

$$R_t \le \mathbb{I}(V; Y). \tag{2}$$

The above lemma gives an achievable region for this problem with strong secrecy (condition 3 of Definition 1). In the following corollary, we show that for every rate pair in this region, there exists a code for which the K-L divergence between the distributions $Q_S Q_{T,Z}$ and $Q_{S,T,Z}$ is small. This property is useful in the proof of Theorem 1, where we show that codes for the above secrecy problem lead to suitable codes and faking procedure for our message deniability problem.

Corollary 1. Lemma 1 continues to hold if the condition $\mathbb{D}(\mathbb{Q}_{S}\mathbb{Q}_{T,\mathbb{Z}} || \mathbb{Q}_{S,T,\mathbb{Z}}) < \delta$ is added to Definition 1.

We discuss the proof of Lemma 1 and Corollary 1 in Appendix A.

B. Proof of achievability in Theorem 1

It suffices to prove the achievability of (R, D) pairs satisfying V - U - X - (Y, Z),

$$0 \le R \le \mathbb{I}(Y; V) + D, \text{ and}$$
$$0 \le D \le \mathbb{I}(U; Y|V) - \mathbb{I}(U; Z|V)$$

Note that such an (R, D) pair may be expressed as $R = R_s + R_t$, and $D = R_t$, where the pair (R_s, R_t) satisfies the inequalities (1) and (2) specified in Lemma 1. The crux of the achievability proof is the following reduction argument. Let $\epsilon, \delta > 0$ be given.

Choose *n* large enough so that there exists a code \mathscr{C} of rate (R_s, R_t) satisfying the achievability of Corollary 1 with the chosen values of ϵ and δ . For the message deniability problem, we decompose the *nR*-length message *m* into two parts – a confidential part *s* of *nR_s* bits, and a leaked part *t* of *nR_t* bits. Next, Alice and Bob encode and decode (s, t) using the code $\mathscr{C} = (ENC, DEC)$. The reliability guarantees for our code thus follow directly from the guarantees on \mathscr{C} proved in Corollary 1. The faking procedure draws *s'* independently at random from the distribution Q_s on $\{0, 1\}^{nR_s}$ and outputs $m^{(r)} = (s', t)$. For the faking procedure thus constructed,

$$\mathbb{D}(\mathbf{Q}_{M^{(\mathrm{F})},\mathbf{Z}} \| \mathbf{Q}_{M,\mathbf{Z}}) = \mathbb{D}(\mathbf{Q}_{S',T,\mathbf{Z}} \| \mathbf{Q}_{S,T,\mathbf{Z}})$$
$$= \mathbb{D}(\mathbf{Q}_{S}\mathbf{Q}_{T,\mathbf{Z}} \| \mathbf{Q}_{S,T,\mathbf{Z}})$$
$$\stackrel{(a)}{\leq} \delta.$$

In the above, the bound (a) follows from the guarantees provided in Corollary 1. This shows that $(R, D) \in \mathscr{R}_m$.

C. Proof of converse in Theorem 1

The scheme described in the previous section has the following property. Given the part of the message that is revealed to Judy, the additional information learnt by Judy based on her channel observation is no larger than δ . In particular this implies that for the scheme presented in our achievability proof, $I(M; \mathbb{Z} | M^{(p)}) < \delta$, *i.e.*, given the fake message, the channel observation and the true message are nearly independent. In our converse proof, we start off by showing that this property must, in fact, be true for any faking procedure that satisfies the plausibility requirement. Further, we also show that in order for a faking procedure to be plausible, the entropy for the message and the fake message must be close each other. The following lemma makes these claims precise.

Lemma 2. Let $M^{(F)}$ be (δ, D) -plausibly deniable for M given observation \mathbb{Z} and satisfy $M^{(F)} - M - \mathbb{Z}$. Then, there exists a non-negative constant λ depending only on $\mathsf{P}_{Z|X}$ and $|\mathcal{M}|$ such that

$$\mathbb{I}(M; \mathbb{Z} | M^{(\mathrm{F})}) \le \delta + n\lambda\sqrt{\delta}, \text{ and} \\ |\mathbb{H}(M) - \mathbb{H}(M^{(\mathrm{F})})| \le \delta + n\lambda\sqrt{\delta}.$$

Proof:

We explicitly prove only the first inequality. The second inequality follow from a similar reasoning. We first use the definition of mutual information and Kullback-Leibler Divergence to note that

$$\begin{split} \mathbb{I}(M; \mathbf{Z}|M^{(v)}) &= \mathbb{H}(\mathbf{Z}|M^{(v)}) - \mathbb{H}(\mathbf{Z}|M) \\ &= \sum_{(\mathbf{z},m): \Omega_{\mathbf{Z},M^{(v)}}(\mathbf{z},m) > 0} \mathbb{Q}_{\mathbf{Z},M^{(v)}}(\mathbf{z},m) \log \frac{\mathbb{Q}_{M^{(v)}}(m)}{\mathbb{Q}_{\mathbf{Z},M^{(v)}}(\mathbf{z},m)} - \sum_{(\mathbf{z},m): \Omega_{\mathbf{Z},M}(\mathbf{z},m) > 0} \mathbb{Q}_{\mathbf{Z},M}(\mathbf{z},m) \log \frac{\mathbb{Q}_{M}(m)}{\mathbb{Q}_{\mathbf{Z},M}(\mathbf{z},m)} \\ &= \sum_{m: \Omega_{M^{(v)}}(m) > 0} \mathbb{Q}_{M^{(v)}}(m) \log \mathbb{Q}_{M^{(v)}}(m) - \sum_{m \in \mathscr{M}} \mathbb{Q}_{M}(m) \log \mathbb{Q}_{M}(m) \\ &- \sum_{(\mathbf{z},m): \Omega_{\mathbf{Z},M^{(v)}}(\mathbf{z},m) > 0} \mathbb{Q}_{\mathbf{Z},M^{(v)}}(\mathbf{z},m) \log \frac{1}{\mathbb{Q}_{\mathbf{Z},M^{(v)}}(\mathbf{z},m)} + \sum_{(\mathbf{z},m): \Omega_{\mathbf{Z},M}(\mathbf{z},m) > 0} \mathbb{Q}_{\mathbf{Z},M}(\mathbf{z},m) \log \frac{1}{\mathbb{Q}_{\mathbf{Z},M}(\mathbf{z},m)} \\ &= D\left(\mathbb{Q}_{M^{(v)}}\|\mathbb{Q}_{M}\right) + \sum_{m \in \mathscr{M}} \left[\mathbb{Q}_{M^{(v)}}(m) - \mathbb{Q}_{M}(m)\right] \log \mathbb{Q}_{M}(m) \\ &- D\left(\mathbb{Q}_{\mathbf{Z},M^{(v)}}\|\mathbb{Q}_{\mathbf{Z},M}\right) - \sum_{(\mathbf{z},m): \Omega_{\mathbf{Z},M^{(v)}}(\mathbf{z},m) > 0} \mathbb{Q}_{\mathbf{Z},M^{(v)}}(\mathbf{z},m) \log \frac{1}{\mathbb{Q}_{\mathbf{Z},M}(\mathbf{z},m)} - \sum_{(\mathbf{z},m): \Omega_{\mathbf{Z},M}(\mathbf{z},m) > 0} \mathbb{Q}_{\mathbf{Z},M}(\mathbf{z},m) \log \frac{1}{\mathbb{Q}_{\mathbf{Z},M}(\mathbf{z},m)} \\ &= D\left(\mathbb{Q}_{M^{(v)}}\|\mathbb{Q}_{\mathbf{Z},M}\right) + \sum_{m \in \mathscr{M}} \left[\mathbb{Q}_{M^{(v)}}(m) - \mathbb{Q}_{M}(m)\right] \log \mathbb{Q}_{M}(m) \\ &- D\left(\mathbb{Q}_{\mathbf{Z},M^{(v)}}\|\mathbb{Q}_{\mathbf{Z},M}\right) + \sum_{m \in \mathscr{M}} \left[\mathbb{Q}_{M^{(v)}}(m) - \mathbb{Q}_{M}(m)\right] \log \mathbb{Q}_{M}(m) \\ &- D\left(\mathbb{Q}_{\mathbf{Z},M^{(v)}}\|\mathbb{Q}_{\mathbf{Z},M}\right) - \sum_{(\mathbf{z},m):\Omega_{\mathbf{Z},M}(\mathbf{z},m) > 0} \left[\mathbb{Q}_{\mathbf{Z},M^{(v)}}(\mathbf{z},m) - \mathbb{Q}_{\mathbf{Z},M}(\mathbf{z},m)\right] \log \frac{1}{\mathbb{Q}_{\mathbf{Z},M}(\mathbf{z},m)}. \end{aligned}$$

$$(3)$$

In the last step we use the fact that $\{(\mathbf{z}, m) : \mathbf{Q}_{\mathbf{Z}, M^{(F)}}(\mathbf{z}, m) > 0\} \subseteq \{(\mathbf{z}, m) : \mathbf{Q}_{\mathbf{Z}, M}(\mathbf{z}, m) > 0\}$ as $\mathbb{D}(\mathbf{Q}_{\mathbf{Z}, M^{(F)}} || \mathbf{Q}_{\mathbf{Z}, M}) < \delta < \infty$. Continuing further from Eq. (3) and again using the fact that $\mathbb{D}(\mathbf{Q}_{\mathbf{Z}, M^{(F)}} || \mathbf{Q}_{\mathbf{Z}, M}) < \delta$, we have

$$\mathbb{I}(M; \mathbf{Z}|M^{(\mathrm{F})}) \stackrel{(a)}{\leq} \delta - \sum_{(\mathbf{z},m): \mathbf{Q}_{\mathbf{Z},M}(\mathbf{z},m)>0} \left[\mathbf{Q}_{\mathbf{Z},M^{(\mathrm{F})}}(\mathbf{z},m) - \mathbf{Q}_{\mathbf{Z},M}(\mathbf{z},m) \right] \log \frac{1}{\mathbf{Q}_{\mathbf{Z},M}(\mathbf{z},m)}$$

$$\begin{split} &\leq \delta + \sum_{(\mathbf{z},m): \Omega_{\mathbf{Z},M}(\mathbf{z},m)>0} \left| \mathsf{Q}_{\mathbf{Z},M^{(\mathrm{F})}}(\mathbf{z},m) - \mathsf{Q}_{\mathbf{Z},M}(\mathbf{z},m) \right| \log \frac{1}{\mathsf{Q}_{\mathbf{Z},M}(\mathbf{z},m)} \\ &= \delta + \sum_{(\mathbf{z},m): \Omega_{\mathbf{Z},M}(\mathbf{z},m)>0} \left| \mathsf{Q}_{\mathbf{Z},M^{(\mathrm{F})}}(\mathbf{z},m) - \mathsf{Q}_{\mathbf{Z},M}(\mathbf{z},m) \right| \log \frac{|\mathcal{M}|}{\sum_{\mathbf{x}:\mathsf{P}_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x})>0} \mathsf{P}_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}) \mathsf{Q}_{\mathbf{X}|M}(\mathbf{x}|m)} \\ &\stackrel{(b)}{\leq} \delta + \sum_{(\mathbf{z},m): \Omega_{\mathbf{Z},M}(\mathbf{z},m)>0} \left| \mathsf{Q}_{\mathbf{Z},M^{(\mathrm{F})}}(\mathbf{z},m) - \mathsf{Q}_{\mathbf{Z},M}(\mathbf{z},m) \right| \sum_{\substack{\mathbf{x}\in\mathcal{Q}^{\mathbb{T}}\\\mathsf{P}_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x})>0}} \mathsf{Q}_{\mathbf{X}|M}(\mathbf{x}|m) \log \frac{|\mathcal{M}|}{\mathsf{P}_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x})} \\ &\leq \delta + \sum_{(\mathbf{z},m): \Omega_{\mathbf{Z},M}(\mathbf{z},m)>0} \left| \mathsf{Q}_{\mathbf{Z},M^{(\mathrm{F})}}(\mathbf{z},m) - \mathsf{Q}_{\mathbf{Z},M}(\mathbf{z},m) \right| \sum_{\substack{\mathbf{x}\in\mathcal{Q}^{\mathbb{T}}\\\mathsf{P}_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x})>0}} \mathsf{Q}_{\mathbf{X}|M}(\mathbf{x}|m) \log \frac{|\mathcal{M}|}{\mathsf{P}_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}')} \\ &\leq \delta + \sum_{(\mathbf{z},m): \Omega_{\mathbf{Z},M}(\mathbf{z},m)>0} \left| \mathsf{Q}_{\mathbf{Z},M^{(\mathrm{F})}}(\mathbf{z},m) - \mathsf{Q}_{\mathbf{Z},M}(\mathbf{z},m) \right| \sum_{\substack{\mathbf{x}\in\mathcal{Q}^{\mathbb{T}}\\\mathsf{P}_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x})>0}} \mathsf{Q}_{\mathbf{X}|M}(\mathbf{x}|m) \log \frac{(|\mathcal{M}|)^{1/n}}{\mathsf{min}_{(\mathcal{Z},\mathcal{X}):\mathsf{P}_{\mathbf{Z}|\mathbf{X}}(\mathcal{Z}|\mathbf{x})} \right| \\ &\leq \delta + n\sqrt{2\delta} \left[\log |\mathcal{M}| - \log \frac{1}{\mathsf{min}_{(\mathcal{Z},\mathcal{X}):\mathsf{P}_{\mathbf{Z}|\mathbf{X}}(\mathcal{Z}|\mathbf{x})>0} \mathsf{P}_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x})} \right]. \end{split}$$

In the above, (*a*) follows by using the fact that $M^{(F)}$ is (δ, D) -plausibly deniable for M given \mathbb{Z} to bound the first term in (3), noting that $Q_M(m)$ equals $1/|\mathscr{M}|$ to conclude that the second term is zero, and applying the non-negativity of the Kullback-Leibler divergence. The inequality (*b*) is obtained by using Jensen's inequality. Finally, (*c*) follows applying Pinsker's inequality to bound the variational distance between the distributions $Q_{\mathbb{Z},\mathscr{M}^{(F)}}$ and $Q_{\mathbb{Z},\mathscr{M}}$.

Proof of converse of Theorem 1:

Let $\epsilon, \delta > 0$. We begin by obtaining *n*-letter bounds on *D* and *R* for any (ϵ, R) -reliable and (δ, D) -plausibly deniable code. To this end, from the definition and Lemma 2, there exists $\gamma = \gamma(\epsilon, \delta) > 0$ such that $\lim_{(\epsilon, \delta) \to (0, 0)} \gamma = 0$, and

$$nD \leq \mathbb{H}(M^{(\mathrm{F})}|M)$$

$$= \mathbb{H}(M|M^{(\mathrm{F})}) + \mathbb{H}(M^{(\mathrm{F})}) - \mathbb{H}(M)$$

$$\leq \mathbb{H}(M|M^{(\mathrm{F})}) + n\gamma$$

$$\stackrel{(a)}{\leq} \mathbb{I}(M; \mathbf{Y}|M^{(\mathrm{F})}) + 2n\gamma$$

$$\leq \mathbb{I}(M; \mathbf{Y}|M^{(\mathrm{F})}) - \mathbb{I}(M; \mathbf{Z}|M^{(\mathrm{F})}) + 3n\gamma.$$
(4)

In the above, (a) follows by applying Fano's inequality and letting γ be at least as large as ϵ . Next, Applying we apply Fano's inequality to bound the rate R as

$$nR \leq \mathbb{I}(M; \mathbf{Y}) + n\gamma$$

= $\mathbb{I}(M^{(F)}, M; \mathbf{Y}) + n\gamma$
= $\mathbb{I}(M^{(F)}; \mathbf{Y}) + \mathbb{I}(M; \mathbf{Y}|M^{(F)}) + n\gamma$
 $\leq \mathbb{I}(M^{(F)}; \mathbf{Y}) + \mathbb{I}(M; \mathbf{Y}|M^{(F)}) - \mathbb{I}(M; \mathbf{Z}|M^{(F)}) + 2n\gamma,$ (5)

where the second equality follows from the fact that $M^{(F)} - M - Y$ is a Markov chain. Next, we obtain single-letter versions of the above expressions. Let *T* be uniformly distributed over [1:n] and independent of $(M, M^{(F)}, \mathbf{X}, \mathbf{Y}, \mathbf{Z})$. From (4),

$$D \leq \frac{1}{n} \left[\mathbb{I}(M; \mathbf{Y} | M^{(\text{F})}) - \mathbb{I}(M; \mathbf{Z} | M^{(\text{F})}) \right] + 3\gamma$$

$$\stackrel{(a)}{=} \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{I}(M; Y_i | Y^{i-1}, Z_{i+1}^n, M^{(\text{F})}) - \mathbb{I}(M; Z_i | Y^{i-1}, Z_{i+1}^n, M^{(\text{F})}) \right] + 3\gamma$$

$$= \mathbb{I}(M; Y_T | Y^{T-1}, Z_{T+1}^n, M^{(\text{F})}, T) - \mathbb{I}(M; Z_T | Y^{T-1}, Z_{T+1}^n, M^{(\text{F})}, T) + 3\gamma$$

where (a) follows from Csiszár's sum identity [16]. Also,

$$\begin{split} \mathbb{I}(M^{\scriptscriptstyle(\mathrm{F})};\mathbf{Y}) &= \sum_{i=1}^{n} \mathbb{I}(M^{\scriptscriptstyle(\mathrm{F})};Y_{i}|Y^{i-1}) \\ &\leq \sum_{i=1}^{n} \mathbb{I}(M^{\scriptscriptstyle(\mathrm{F})},Y^{i-1},Z_{i+1}^{n};Y_{i}) \\ &= n\mathbb{I}(M^{\scriptscriptstyle(\mathrm{F})},Y^{T-1},Z_{T+1}^{n};Y_{T}|T) \\ &\leq n\mathbb{I}(M^{\scriptscriptstyle(\mathrm{F})},Y^{T-1},Z_{T+1}^{n},T;Y_{T}). \end{split}$$

Hence, from (5),

v

$$R \leq \mathbb{I}(M^{(\text{F})}, Y^{T-1}, Z^{n}_{T+1}, T; Y_{T}) + \mathbb{I}(M; Y_{T}|Y^{T-1}, Z^{n}_{T+1}, M^{(\text{F})}, T) - \mathbb{I}(M; Z_{T}|Y^{T-1}, Z^{n}_{T+1}, M^{(\text{F})}) + 2\gamma.$$

Next, let $V = (M^{(F)}, Y^{T-1}, Z_{T+1}^n, T)$, U = (V, M), $X = X_T$, $Y = Y_T$ and $Z = Z_T$. Then, clearly, V - U - X - (Y, Z). Substituting above and letting ϵ and δ be arbitrarily small (but positive) shows that any achievable rate-deniability pair (R, D) must satisfy

$$0 \le R \le \mathbb{I}(Y; V) + \mathbb{I}(U; Y|V) - \mathbb{I}(U; Z|V), \text{ and} \\ 0 \le D \le \min \{R, \mathbb{I}(U; Y|V) - \mathbb{I}(U; Z|V)\}$$

for some random variables U and V satisfying the Markov chain V - U - X - (Y, Z).

Finally, we argue that it suffices to consider random variables U and V such that $|\mathcal{U}| \le (|\mathcal{X}| + 1)(|\mathcal{X}| + 2)$ and $|\mathcal{V}| \le |\mathcal{X}| + 2$. The proof follows along the cardinality bounding argument for the broadcast channel with confidential messages [3, pp. 347-348]. In particular, consider auxiliary variables V and U, that take values in sets \mathscr{V} and \mathscr{U} respectively, and are jointly distributed with $X, Y, \text{ and } Z \text{ such that } \mathsf{P}_{VUXYZ}(v, u, x, y, z) = \mathsf{P}_{V}(v) \mathsf{P}_{U|V}(u|v) \mathsf{P}_{X|U}(x|u) \mathsf{P}_{YZ|X}(y, z|x) \text{ for every } (v, u, x, y, z) \in \mathcal{V} \times \mathcal{U} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}.$ The first step in the proof is to show that there exist auxiliary variables \tilde{V} and \tilde{U} , that take values in sets $\tilde{\mathscr{V}}$ and \mathscr{U} respectively, are jointly distributed with X, Y, and Z such that $\mathsf{P}_{\tilde{V}\tilde{U}XYZ}(v, u, x, y, z) = \mathsf{P}_{\tilde{V}}(v)\mathsf{P}_{U|V}(u|v)\mathsf{P}_{X|U}(x|u)\mathsf{P}_{YZ|X}(y, z|x)$ for every $(v, u, x, y, z) \in \tilde{\mathcal{V}} \times \mathcal{U} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, where, $\mathsf{P}_{\tilde{V}}$ satisfy the following constraints:

$$\sum_{v \in \tilde{\mathcal{V}}} \mathsf{P}_{\tilde{V}}(v) \sum_{u \in \mathcal{U}} \mathsf{P}_{U|V}(u|v) \mathsf{P}_{X|U}(x|u) = \sum_{v \in \mathcal{V}} \mathsf{P}_{V}(v) \sum_{u \in \mathcal{U}} \mathsf{P}_{U|V}(u|v) \mathsf{P}_{X|U}(x|u) = \mathsf{P}_{X}(x) \text{ for all } x \in \mathscr{X},$$
(6)

$$\sum_{\nu \in \tilde{\mathcal{V}}} \mathsf{P}_{\tilde{\mathcal{V}}}(\nu) \mathbb{H}(Y|V=\nu) = \sum_{\nu \in \tilde{\mathcal{V}}} \mathsf{P}_{V}(\nu) \mathbb{H}(Y|V=\nu), \tag{7}$$

$$\sum_{v\in\tilde{\mathscr{V}}}\mathsf{P}_{\tilde{V}}(v)\bigg(\mathsf{H}(Y|V=v)-\sum_{u\in\mathscr{U}}\mathsf{P}_{U|V}(u|v)\mathsf{H}(Y|U=u)\bigg)=\sum_{v\in\mathscr{V}}\mathsf{P}_{V}(v)\bigg(\mathsf{H}(Y|V=v)-\sum_{u\in\mathscr{U}}\mathsf{P}_{U|V}(u|v)\mathsf{H}(Y|U=u)\bigg),\tag{8}$$

$$\sum_{v \in \tilde{\mathcal{V}}} \mathsf{P}_{\tilde{\mathcal{V}}}(v) \left(\mathbb{H}(Z|V=v) - \sum_{u \in \mathcal{U}} \mathsf{P}_{U|V}(u|v) \mathbb{H}(Z|U=u) \right) = \sum_{v \in \mathcal{V}} \mathsf{P}_{V}(v) \left(\mathbb{H}(Z|V=v) - \sum_{u \in \mathcal{U}} \mathsf{P}_{U|V}(u|v) \mathbb{H}(Z|U=u) \right), \text{ and}$$
(9)
$$|\tilde{\mathcal{V}}| \le |\mathcal{X}| + 2.$$
(10)

In the above, constraints. (6) and (7) ensure that $\mathbb{I}(Y; \tilde{V})$ equals $\mathbb{I}(Y; V)$, (8) and (9) ensure that $\mathbb{I}(\tilde{U}; Y|\tilde{V}) - \mathbb{I}(\tilde{U}; Z|\tilde{V})$ equals $\mathbb{I}(U; Y|V) - \mathbb{I}(U; Z|V)$, and Eq. (10) follows from Caratheodory's theorem (c.f. [17, Lemma 3]) as Eqs. (6)-(9) imply at most $|\mathscr{X}| + 2$ constraints on $\mathsf{P}_{\tilde{V}}$. Note that the number of constraints in our setting is one less than that in [3] as we do not require $\mathbb{I}(Z; \tilde{V})$ to equal $\mathbb{I}(Z; V)$. Next, using a similar reasoning, the next step is to show that there exists an auxiliary variable \hat{U} that takes values in a set $\hat{\mathcal{U}}$, is jointly distributed with \tilde{V}, X, Y , and Z such that $\mathsf{P}_{\tilde{V}\hat{U}XYZ}(v, u, x, y, z) = \mathsf{P}_{\tilde{V}}(v)\mathsf{P}_{\hat{U}|\tilde{V}}(u|v)\mathsf{P}_{X|U}(x|u)\mathsf{P}_{YZ|X}(y, z|x)$ for every $(v, u, x, y, z) \in \tilde{\mathcal{V}} \times \hat{\mathcal{U}} \times \hat{\mathcal{X}} \times \hat{\mathcal{Y}} \times \hat{\mathcal{Z}}$, where, for each $v \in \tilde{\mathcal{V}}$, $\mathsf{P}_{\hat{\mathcal{U}}|\tilde{\mathcal{V}}}$ satisfies the following constraints:

$$\sum_{u \in \hat{\mathscr{U}}} \mathsf{P}_{\hat{U}|\tilde{V}}(u|v) \mathsf{P}_{X|U}(x|u) = \sum_{u \in \hat{\mathscr{U}}} \mathsf{P}_{U|\tilde{V}}(u|v) \mathsf{P}_{X|U}(x|u) = \mathsf{P}_{X|\tilde{V}}(x|v),$$
(11)

$$\sum_{u \in \mathscr{U}} \mathsf{P}_{\hat{U}|\tilde{V}}(u|v) \mathbb{H}(Y|U=u) = \sum_{u \in \mathscr{U}} \mathsf{P}_{U|\tilde{V}}(u|v) \mathbb{H}(Y|U=u),$$
(12)

$$\sum_{u \in \mathscr{U}}^{u \in \mathscr{U}} \mathsf{P}_{\hat{U}|\tilde{V}}(u|v) \mathbb{H}(Z|U=u) = \sum_{u \in \mathscr{U}}^{u \in \mathscr{U}} \mathsf{P}_{U|\tilde{V}}(u|v) \mathbb{H}(Z|U=u), \text{ and}$$
(13)

$$\left|\left\{u \in \widehat{\mathscr{U}} : \mathsf{P}_{\widehat{U}|\widehat{V}}(u|v) > 0\right\}\right| \le |\mathscr{X}| + 1.$$
(14)

Here, constraint (11) ensures consistency of the marginals of $\mathsf{P}_{\tilde{V}\hat{U}XYZ}$ and $\mathsf{P}_{\tilde{V}\hat{U}XYZ}$ with respect to (\tilde{V}, X, Y, Z), constraints (12) and (13) (along with (11)) ensure that $\mathbb{I}(\hat{U}; Y|\tilde{V}) - \mathbb{I}(\hat{U}; Z|\tilde{V})$ equals $\mathbb{I}(\tilde{U}; Y|\tilde{V}) - \mathbb{I}(\tilde{U}; Z|\tilde{V})$, and Eq. (14) again follows from Caratheodory's theorem as Eqs. (11)-(13) imply at most $|\mathscr{X}| + 1$ constraints on $\mathsf{P}_{\hat{U}|\tilde{V}}(\cdot|v)$. Finally, summing the bound from (14) over all $v \in \tilde{\mathcal{V}}$, we obtain that it suffices to let $|\hat{\mathcal{U}}|$ be at most $(|\mathcal{X}| + 1)(|\mathcal{X}| + 2)$. This completes the proof of the converse.

D. Discussions

1) Plausible deniability vs Secrecy: In the following discussion, we compare the capacity region \mathscr{R}_m to rate regions for two standard information-theoretic secrecy problems - the Wire-Tap Channel [2] and Broadcast Channel with Confidential messages [3] (see Figure 4). To this end, we first adapt the following definitions from [2], [3].

Definition 2 (Rate-Equivocation Region). For a channel $P_{YZ|X}$, the rate-equivocation region \mathcal{R}_{equiv} is the set of all non-negative (R, R_e) pairs such for any $\epsilon > 0$ and large enough block-length *n*, there exists a code for the Wire-Tap Channel problem (Figure 4a) when the message $|\mathcal{M}| \ge 2^{nR}$, $Q_{M(m)} = 1/|\mathcal{M}|$ for each $m \in \mathcal{M}$, $Q_{M,\mathbf{X},\mathbf{Y}}(m \neq \hat{m}) < \epsilon$, and $\mathbb{H}(M|\mathbf{Z}) \ge nR_e$.



(b) Broadcast Channel with Confidential Messages

Fig. 4: In the Wire-Tap Channel problem (first introduced by [2] and explored further in [3]), the goal for Alice is to transmit a confidential message *m* to the legitimate receiver Bob while ensuring that the "leakage" to the eavesdropper Judy (measured through the rate of equivocation) is smaller than a threshold. The capacity region for this problem (see Definition 2 exhibits a tradeoff between the message rate *R* and the equivocation rate R_e . The Broadcast Channel with Confidential messages setup (introduced by [3]) generalizes the Wire-Tap Channel model to include a "public" message m_0 that is meant to be decoded by both Bob and Judy. Similarly to the Wire-Tap Channel, this setup also includes a confidential message m_1 that is meant to be decoded by only Bob while ensuring that the leakage to Judy is smaller than a threshold. In general, the capacity region for this setup exhibits a tradeoff between three parameters – the rate of the public message R_0 , the rate of the confidential message R_1 , and the equivocation rate. In our discussion, we only consider a two-dimensional projection of this region (see Definition 3) to the set of (R_0, R_1) pairs that ensure that the equivocation about the message m_1 is arbitrarily close to the entropy of m_1 . The reader is referred to [4] for an excellent introduction to these and other information-theoretic security problems.

Definition 3 (Sum Capacity with Confidential and Public messages). For a channel $P_{Y,Z|X}$, the sum capacity region with confidential and public messages \mathscr{R}_{bcc} is the set of all non-negative (R, R_1) pairs with $R \ge R_1$ for which, given any $\epsilon > 0$, for a large enough blocklength *n*, there exists a code for the Broadcast Channel with Confidential Messages setup (Figure 4b) with $|\mathscr{M}_0| \ge 2^{n(R-R_1)}$, $|\mathscr{M}_1| \ge 2^{nR_1}$, $Q_{M_0,M_1}(m_0, m_1) = 1/|\mathscr{M}_0||\mathscr{M}_1|$ for each $(m_0, m_1) \in \mathscr{M}_0 \times \mathscr{M}_1$, $Q_{M_0,M_1,\mathbf{X},\mathbf{Y},\mathbf{Z}}\left((\hat{\mathcal{M}}_0, \hat{\mathcal{M}}_0, \hat{\mathcal{M}}_1) \ne (\mathcal{M}_0, \mathcal{M}_0, \mathcal{M}_1)\right) < \epsilon$ and $\mathbb{H}(\mathcal{M}_1|\mathbf{Z}) \ge nR_1 - \epsilon$.

We note that in the Message Deniability setting, the existence of (δ, D) -plausibly deniable faking procedure implies that the

equivocation of M given Z is no smaller than $D - O(\sqrt{\delta})$.

Proposition 1. Let $M^{(F)}$ be (δ, D) -plausibly deniable for M given observation \mathbb{Z} and satisfy $M^{(F)} - M - \mathbb{Z}$. Then, there exists μ depending only on $P_{Z|X}$ such that

$$\mathbb{H}(M|\mathbf{Z}) \ge nD - n\mu\sqrt{\delta} - 2\delta.$$

Proof:

The above proposition is a direct consequence of Lemma 2. Specifically, note that there exists $\lambda = \lambda(P_{Z|X})$ such that

$$\begin{split} \mathbb{H}(M|\mathbf{Z}) &\geq \mathbb{H}(M|\mathbf{Z}) + \mathbb{I}(M;\mathbf{Z}|M^{(\mathrm{F})}) - \delta - n\lambda\sqrt{\delta} \\ &= \mathbb{H}(M|\mathbf{Z}) + \mathbb{H}(M|M^{(\mathrm{F})}) - \mathbb{H}(M|\mathbf{Z}, M^{(\mathrm{F})}) - \delta - n\lambda\sqrt{\delta} \\ &\geq \mathbb{H}(M|M^{(\mathrm{F})}) - \delta - n\lambda\sqrt{\delta} \\ &= \mathbb{H}(M^{(\mathrm{F})}|M) + \mathbb{H}(M) - \mathbb{H}(M^{(\mathrm{F})}) - \delta - n\lambda\sqrt{\delta} \\ &\geq nD - 2\delta - 2n\lambda\sqrt{\delta}. \end{split}$$

The above proposition leads to the following corollary.

Corollary 2. $\mathscr{R}_{bcc} \subseteq \mathscr{R}_m \subseteq \mathscr{R}_{equiv}$.

Proof:

As proved in [3], \mathscr{R}_{bcc} is the set of all (R, R_1) pairs such that there exist random variables V and U satisfying V - U - X - (Y, Z) and

$$0 \le R \le \min\{\mathbb{I}(V; Y), \mathbb{I}(V; Z)\} + \mathbb{I}(U; Y|V) - \mathbb{I}(U; Z|V)$$
$$0 \le R_0 \le \min\{\mathbb{I}(V; Y), \mathbb{I}(V; Z)\}$$

The first inclusion, $\mathscr{R}_{bcc} \subseteq \mathscr{R}_m$, follows directly by comparing our characterization of \mathscr{R}_m with the above capacity expression. Note that in the setting of [3], the public message of rate R_0 is intended to be decoded by both the receivers, while in our achievability proof of Theorem 1, we require that it be decoded only by Bob. This allows us to operate with public message rates as high as $\mathbb{I}(V; Y)$, rather than min { $\mathbb{I}(V; Y)$, $\mathbb{I}(V; Z)$ } as in [3]. Next, applying Proposition 1 to a sequence of codes with δ approaching zero, we obtain that every $(R, D) \in \mathscr{R}_m$ also lies in \mathscr{R}_{equiv} .

The following example illustrates that both inclusions in the above corollary may be strict.

Example 2 (Binary Erasure Eavesdropper). Consider the example of Figure 1. Let $\mathscr{X} = \mathscr{Y} = \{0, 1\}, \ \mathscr{Z} = \{0, \bot, 1\}$, and

$$\mathsf{P}_{YZ|X}(yz|x) = \begin{cases} 1-p & \text{if } (y,z) = (x,x), \\ p & \text{if } (y,z) = (x,\perp), \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

As this is a degraded channel, it suffices to let the variable U in Theorem 1 be equal to X. Further, using the fact that Y = Xand H(X|Z, V) = pH(X|V) (as the channel from X to Z is a Binary Erasure Channel with erasure probability p), we obtain the following characterisation for \mathscr{R}_m . \mathscr{R}_m is the set of (R, D) pairs such that there exists a random variable V with V - X - Z,

$$0 \le R \le \mathbb{H}(X) - (1 - p)\mathbb{H}(X|V), \text{ and}$$
$$0 \le D \le \min\{R, p\mathbb{H}(X|V)\}.$$

Let $\alpha_{X,V} = H(X|V)/H(X)$. Thus, $\mathbb{H}(X) - (1-p)\mathbb{H}(X|V) = (1 - \alpha_{X,V}(1-p))\mathbb{H}(X)$, and $p\mathbb{H}(X) = p\alpha_{X,V}\mathbb{H}(X)$. Note that $\alpha_{X,V}$ may take any value in the interval [0, 1] and the maximum value of $\mathbb{H}(X)$ equals 1. Thus, \mathscr{R}_m consists of (R, D) pairs such that for some $\alpha \in [0, 1]$, $0 \le R \le (1 - \alpha(1 - p))$ and $0 \le D \le \min\{R, \alpha p\}$. Simplifying further, we conclude that the region \mathscr{R}_m consists of (R, D) pairs such that

$$0 \le R \le 1$$

$$0 \le D \le \min\left\{\frac{p(1-R)}{1-p}, R\right\}$$

We next compare this region with the regions \mathscr{R}_{bcc} and \mathscr{R}_{equiv} . For the channel considered in this example, the Rate-Equivocation region consists of all (R, R_e) pairs satisfying

$$0 \le R \le 1$$

$$0 \le R_e \le \min\{p, R\}.$$

$$0 \le R \le 1$$

 $0 \le R_1 \le \min\left\{\frac{p(1-p-R)}{1-2p}, R\right\}.$

Comparing the above regions, it is evident that the inclusion relation in Corollary 2 may be strict. The plot shown in Figure 5 compares these regions.



Fig. 5: Comparision of \mathscr{R}_m with \mathscr{R}_{bcc} and \mathscr{R}_{equiv} in Example 2.

2) *Rate of deniability as the Equivocation rate:* Even though we define the rate of deniability as an operational property of the faking procedure, surprisingly, it also has a rough interpretation as the rate of equivocation given the eavesdropper channel output as well as the fake message. This is especially interesting in light of Example 2 that shows that the rate of deniability may be strictly smaller than the equivocation rate at the eavesdropper in the Wire-Tap Channel setting. The following proposition states this property formally.

Proposition 2. Let $M^{(F)}$ be (δ, D) -plausibly deniable for M given observation \mathbb{Z} and satisfy $M^{(F)} - M - \mathbb{Z}$. Then, there exists $\mu \ge 0$ depending only on $\mathsf{P}_{YZ|X}$ such that

$$nD - \delta - n\mu\sqrt{\delta} \le \mathbb{H}(M|M^{(F)}, \mathbb{Z}) \le nD + \delta + n\mu\sqrt{\delta}.$$

Proof: Note that

$$\mathbb{H}(M|M^{(\mathrm{F})}, \mathbf{Z}) = \mathbb{H}(M|M^{(\mathrm{F})}) - \mathbb{I}(M; \mathbf{Z}|M^{(\mathrm{F})})$$

= $\mathbb{H}(M^{(\mathrm{F})}|M) - \mathbb{H}(M^{(\mathrm{F})}) + \mathbb{H}(M) - \mathbb{I}(M; \mathbf{Z}|M^{(\mathrm{F})})$
= $nD - \mathbb{H}(M^{(\mathrm{F})}) + \mathbb{H}(M) - \mathbb{I}(M; \mathbf{Z}|M^{(\mathrm{F})}).$

Applying Lemma 2 and the non-negativity of mutual information to the terms on the left hand side above gives the claimed result.

V. TRANSMITTER AND RECEIVER DENIABILITY

Before formally proving Theorems 2 and 3, we introduce the notion of *zero information variables* that is central to our discussion of the achievability proofs presented in this section.

A. Zero Information Variables

For a random variable $W \sim \mathsf{P}_W$ and a channel $\mathsf{P}_{Z|W}$, we define the following relation: for $w_1, w_2 \in \mathcal{W}$, we say that $w_1 \sim w_2$ if $\mathsf{P}_{Z|W}(z|w_1) = \mathsf{P}_{Z|W}(z|w_2)$, for all $z \in \mathscr{Z}$. It is evident that this is an equivalence relation. Let \mathscr{U}_0 represent the set of equivalence classes of this relation. We define the *zero-information* random variable U_0 of W w.r.t. $\mathsf{P}_{Z|W}$ as a random variable taking values in \mathscr{U}_0 and jointly distributed with W and Z such that $W \in U_0$ with probability 1. For each $w \in \mathcal{W}$, we will call the corresponding u_0 its *zero-information symbol*.

Note that, U_0 is a function of W. Intuitively, the zero information symbol u_0 of w is the largest subset of \mathcal{W} such that each $w' \in u_o$ is statistically indistinguishable from w given any $z \in \mathscr{Z}$ with $\mathsf{P}_{Z|W}(z|w) > 0$. Figure 6 shows an example



Fig. 6: Let *W* and *Z* be random variables distributed on $\mathscr{W} = \{w_1, w_2, w_3\}$ and $\mathscr{Z} = \{z_1, z_2, z_3\}$ respectively with $\mathsf{P}_{Z|W}$ specified according to the edge labels in the above figure. Notice that w_1 and w_2 are indistinguishable to an observer who has access to only *Z* as $\mathsf{P}_{Z|W}(z|w_1) = \mathsf{P}_{Z|W}(z|w_2)$ for every $z \in \mathscr{Z}$. Hence, the zero-information variable for the distribution $\mathsf{P}_{Z|W}$ takes the values $u_{0,1} \equiv \{w_1, w_2\}$ and $u_{0,2} \equiv \{w_3\}$.

of a zero-information variable. Note that $U_0 - W - Z$ (since U_0 is a function of W), $W - U_0 - Z$ (by definition), and $\mathsf{P}_{Z|W}(z|w) = \mathsf{P}_{Z|W,U_0}(z|w, u_o) = \mathsf{P}_{Z|U_0}(z|u_o)$ if u_o is the zero-information symbol of w.

In our achievability proofs for Transmitter and Receiver deniability, the use of zero-information variables considerably simplifies the proof. In particular, we argue that the rate regions claimed achievable in Theorems 2 and 3 it suffices to consider zero information variables instead of the general class of auxiliary variables presented in the theorem statements. The following lemma shows that such a choice does not lead to any loss of optimality.

Lemma 3. Suppose W - U - Z and U - W - (V,Z) are Markov chains. Then $\mathbb{I}(W; V|U) \leq \mathbb{I}(W; V|U_0)$, where U_0 is the zero-information random variable of W w.r.t. $\mathsf{P}_{Z|W}$.

Proof:

We first show that the Markov chains W - U - Z and U - W - Z imply that U_0 must also be a function of U. To show this, it is enough to show that for $w_1, w_2 \in \mathcal{W}$ with $\mathsf{P}_W(w_1), \mathsf{P}_W(w_2) > 0$, if there is a $z \in \mathscr{Z}$ such that $\mathsf{P}_{Z|W}(z|w_1) \neq \mathsf{P}_{Z|W}(z|w_2)$, then for every $u \in \mathscr{U}$ at least one of $\mathsf{P}_{U|W}(u|w_1)$ and $\mathsf{P}_{U|W}(u|w_2)$ must be zero. Suppose, to the contrary both $\mathsf{P}_{U|W}(u|w_1), \mathsf{P}_{U|W}(u|w_2) > 0$. Then

$$\mathsf{P}_{Z|W}(z|w_1) \stackrel{(a)}{=} \mathsf{P}_{Z|W,U}(z|w_1, u)$$

$$\stackrel{(b)}{=} \mathsf{P}_{Z|U}(z|u)$$

$$\stackrel{(c)}{=} \mathsf{P}_{Z|W,U}(z|w_2, u)$$

$$\stackrel{(d)}{=} \mathsf{P}_{Z|W}(z|w_2),$$

where (a) follows from the Markov chain U - V - Z and the fact that $\mathsf{P}_W(w_1)\mathsf{P}_{U|W}(u|w_1) > 0$; (b) follows from the Markov chain W - U - Z; (c) follows from the Markov chain W - U - Z and the fact that $\mathsf{P}_W(w_2)\mathsf{P}_{U|W}(u|w_2) > 0$; and (d) follows from the Markov chain U - W - Z. But, this is a contradiction.

Thus, U_0 is a function of U. From its definition, U_0 is a function of W. Hence,

$$\begin{split} \mathbb{I}(W; V|U) &= \mathbb{I}(W; V|U, U_0) \\ &\leq \mathbb{I}(U, W; V|U_0) \\ &= \mathbb{I}(W; V|U_0) + \mathbb{I}(U; V|W, U_0) \\ &\stackrel{(a)}{=} \mathbb{I}(W; V|U_0) + \mathbb{I}(U; V|W) \\ &\stackrel{(b)}{=} \mathbb{I}(W; V|U_0). \end{split}$$

where (a) uses the fact that U_0 is a function of W and (b) follows from U - W - V being a Markov chain.

B. Transmitter Deniability

 $\pi(\mathbf{X}, \mathbf{7}) \times (\mathbf{F}) = \pi \mathbf{1}(\mathbf{7}) \times (\mathbf{F}) = \pi \mathbf{1}(\mathbf{7}) \times (\mathbf{F})$

We begin our proof for Theorem 2 by stating two lemmas that lead to our converse arguments. The following lemma mirrors Lemma 2 from the message deniability setting and derives necessary conditions for any faking procedure to be plausible with respect to the eavesdropper's observation. In particular, we show that for any plausibly deniable faking procedure, the true codeword and the eavesdropper observation must be nearly conditionally independent given the fake codeword. Further, the joint distribution of the true and fake codewords must be such that it allows exchanging *M* for $M^{(F)}$ (and *vice versa*) does not changes entropic terms involving these by at most δ .

Lemma 4. Let $\mathbf{X}^{(F)}$ be (δ, D) -plausibly deniable for \mathbf{X} given observation \mathbf{Z} and satisfy $\mathbf{X}^{(F)} - \mathbf{X} - \mathbf{Z}$. Then, there exists a constant κ depending only on $\mathsf{P}_{Z|X}$ such that

$$\begin{split} \mathbb{I}(\mathbf{X}; \mathbf{Z} | \mathbf{X}^{(\mathrm{F})}) &\leq n\kappa\sqrt{\delta}, \\ \left| \mathbb{H}(\mathbf{X} | \mathbf{X}^{(\mathrm{F})}) - \mathbb{H}(\mathbf{X}^{(\mathrm{F})} | \mathbf{X}) \right| &\leq n\kappa\sqrt{\delta}, \\ \left| \mathbb{H}(\mathbf{X}) - \mathbb{H}(\mathbf{X}^{(\mathrm{F})}) \right| &\leq n\kappa\sqrt{\delta}, \text{ and} \\ \left| \mathbb{H}(\mathbf{X} | \mathbf{X}^{(\mathrm{F})}, M) - \mathbb{H}(\mathbf{X}^{(\mathrm{F})} | \mathbf{X}, \mathrm{MSG}(\mathbf{X}^{(\mathrm{F})})) \right| &\leq n\kappa\sqrt{\delta}. \end{split}$$

Proof:

We explicitly only prove the first inequality. The other inequalities follow from a similar reasoning.

$$\begin{split} & \| (\mathbf{X}; \mathbf{Z} | \mathbf{X}^{(r)}) = \| (\mathbf{Z} | \mathbf{X}^{(r)}) - \| (\mathbf{Z} | \mathbf{X}) \\ &= \sum_{(\mathbf{z}, \mathbf{x}): \mathbf{Q}_{\mathbf{Z}, \mathbf{X}^{(r)}}(\mathbf{z}, \mathbf{x}) \log \frac{\mathbf{Q}_{\mathbf{X}^{(r)}}(\mathbf{x})}{\mathbf{Q}_{\mathbf{Z}, \mathbf{X}^{(r)}}(\mathbf{z}, \mathbf{x})} - \sum_{(\mathbf{z}, \mathbf{x}): \mathbf{Q}_{\mathbf{Z}, \mathbf{X}}(\mathbf{z}, \mathbf{x}) > 0} \mathbf{Q}_{\mathbf{Z}, \mathbf{X}}(\mathbf{z}, \mathbf{x}) \log \frac{\mathbf{Q}_{\mathbf{X}}(\mathbf{x})}{\mathbf{Q}_{\mathbf{Z}, \mathbf{X}}(\mathbf{z}, \mathbf{x})} \\ &= D \left(\mathbf{Q}_{\mathbf{X}^{(r)}} \| \| \mathbf{Q}_{\mathbf{X}} \right) - D \left(\mathbf{Q}_{\mathbf{Z}, \mathbf{X}^{(r)}} \| \| \mathbf{Q}_{\mathbf{Z}, \mathbf{X}} \right) + \sum_{(\mathbf{z}, \mathbf{x}): \mathbf{Q}_{\mathbf{Z}^{(r)}, \mathbf{X}}(\mathbf{z}, \mathbf{x}) > 0} \mathbf{Q}_{\mathbf{Z}, \mathbf{X}^{(r)}}(\mathbf{z}, \mathbf{x}) \log \frac{\mathbf{Q}_{\mathbf{X}}(\mathbf{x})}{\mathbf{Q}_{\mathbf{Z}, \mathbf{X}}(\mathbf{z}, \mathbf{x})} - \sum_{(\mathbf{z}, \mathbf{x}): \mathbf{Q}_{\mathbf{Z}, \mathbf{X}}(\mathbf{z}, \mathbf{x}) > 0} \mathbf{Q}_{\mathbf{Z}, \mathbf{X}^{(r)}}(\mathbf{z}, \mathbf{x}) \log \frac{\mathbf{Q}_{\mathbf{X}}(\mathbf{x})}{\mathbf{Q}_{\mathbf{Z}, \mathbf{X}}(\mathbf{z}, \mathbf{x})} - \sum_{(\mathbf{z}, \mathbf{x}): \mathbf{Q}_{\mathbf{Z}, \mathbf{X}}(\mathbf{z}, \mathbf{x}) > 0} \mathbf{Q}_{\mathbf{Z}, \mathbf{X}^{(r)}}(\mathbf{z}, \mathbf{x}) \log \frac{\mathbf{Q}_{\mathbf{X}}(\mathbf{x})}{\mathbf{Q}_{\mathbf{Z}, \mathbf{X}}(\mathbf{z}, \mathbf{x})} - \sum_{(\mathbf{z}, \mathbf{x}): \mathbf{Q}_{\mathbf{Z}, \mathbf{X}}(\mathbf{z}, \mathbf{x}) > 0} \mathbf{Q}_{\mathbf{Z}, \mathbf{X}^{(r)}}(\mathbf{z}, \mathbf{x}) \log \frac{\mathbf{Q}_{\mathbf{X}}(\mathbf{x})}{\mathbf{Q}_{\mathbf{Z}, \mathbf{X}}(\mathbf{z}, \mathbf{x})} - \sum_{(\mathbf{z}, \mathbf{x}): \mathbf{Q}_{\mathbf{Z}, \mathbf{X}}(\mathbf{z}, \mathbf{x}) > 0} \mathbf{Q}_{\mathbf{Z}, \mathbf{X}^{(r)}}(\mathbf{z}, \mathbf{x}) \log \frac{\mathbf{Q}_{\mathbf{X}}(\mathbf{x})}{\mathbf{Q}_{\mathbf{Z}, \mathbf{X}}(\mathbf{z}, \mathbf{x}) \log \mathbf{Q}_{\mathbf{Z}, \mathbf{X}}(\mathbf{z}, \mathbf{x})} \log \frac{\mathbf{Q}_{\mathbf{X}}(\mathbf{x})}{\mathbf{Q}_{\mathbf{Z}, \mathbf{X}}(\mathbf{z}, \mathbf{x})} \log \frac{\mathbf{Q}_{\mathbf{X}}(\mathbf{x})}{\mathbf{Q}_{\mathbf{X}}(\mathbf{$$

In the above, step (*a*) uses the fact that $\{(\mathbf{z}, \mathbf{x}) : Q_{\mathbf{Z}^{(F)},\mathbf{X}}(\mathbf{z}, \mathbf{x}) > 0\} \subseteq \{(\mathbf{z}, \mathbf{x}) : Q_{\mathbf{Z},\mathbf{X}}(\mathbf{z}, \mathbf{x}) > 0\}$ (as $D(Q_{\mathbf{Z},\mathbf{X}^{(F)}} || Q_{\mathbf{Z},\mathbf{X}}) < \delta < \infty$). In step (*b*), we use the fact that $\mathbf{X}^{(F)}$ is (δ, D) -plausibly deniable \mathbf{X} for \mathbf{Z} . The bound on the first term follow from definition, the second from non-negativity of K-L divergence, while the last term is bounded by applying Pinsker's inequality.

The following lemma follows from a standard chain of information inequalities with Lemma 4 as a starting point and single-letterizing the resulting expressions.

Lemma 5. Let \mathscr{C} be an (ϵ, R) -reliable code of blocklength *n* for a channel $\mathsf{P}_{YZ|X}$, and let $\mathbf{X}^{(\mathsf{F})}$ be (δ, D) -plausibly deniable for **X** given observation **Z** and satisfy $\mathbf{X}^{(\mathsf{F})} - \mathbf{X} - \mathbf{Z}$. Then, there exists random variables U, X, Y, and Z satisfying U - X - (Y, Z) and a constant $\gamma = \gamma(\epsilon, \delta) > 0$ satisfying $\lim_{(\epsilon, \delta) \to (0, 0)} \gamma = 0$ such that

$$\begin{split} &R \leq \mathbb{I}(X;Y) + \gamma, \\ &D \leq \mathbb{I}(X;Y|U) + \gamma, \text{ and} \\ &\mathbb{I}(X;Z|U) \leq \gamma. \end{split}$$

Proof:

Note that $\mathbf{Y} - \mathbf{X} - \mathbf{X}^{(F)}$. We use Lemma 4 below.

$$\begin{split} nD &\leq \mathbb{H}(\mathrm{MSG}(\mathbf{X}^{(\mathrm{F})})|\mathbf{X}) \\ &= \mathbb{H}(\mathbf{X}^{(\mathrm{F})}|\mathbf{X}) - \mathbb{H}(\mathbf{X}^{(\mathrm{F})}|\mathbf{X}, \mathrm{MSG}(\mathbf{X}^{(\mathrm{F})})) \\ &\stackrel{(a)}{\leq} \mathbb{H}(\mathbf{X}|\mathbf{X}^{(\mathrm{F})}) - \mathbb{H}(\mathbf{X}|\mathbf{X}^{(\mathrm{F})}, \mathrm{MSG}(\mathbf{X})) + 2n\kappa\sqrt{\delta} \\ &\leq \mathbb{H}(\mathbf{X}|\mathbf{X}^{(\mathrm{F})}) - \mathbb{H}(\mathbf{X}|\mathbf{Y}, \mathbf{X}^{(\mathrm{F})}, \mathrm{MSG}(\mathbf{X})) + 2n\kappa\sqrt{\delta} \\ &= \mathbb{H}(\mathbf{X}|\mathbf{X}^{(\mathrm{F})}) - \mathbb{H}(\mathbf{X}|\mathbf{Y}, \mathbf{X}^{(\mathrm{F})}) + \mathbb{I}(\mathbf{X}; \mathrm{MSG}(\mathbf{X})|\mathbf{Y}, \mathbf{X}^{(\mathrm{F})}) + 2n\kappa\sqrt{\delta} \\ &\stackrel{(b)}{\leq} \mathbb{I}(\mathbf{X}; \mathbf{Y}|\mathbf{X}^{(\mathrm{F})}) + n\epsilon + 2n\kappa\sqrt{\delta} \end{split}$$

$$\begin{split} &\overset{(c)}{\leq} \mathbb{I}(\mathbf{X};\mathbf{Y}|\mathbf{X}^{(\text{F})}) - \mathbb{I}(\mathbf{X};\mathbf{Z}|\mathbf{X}^{(\text{F})}) + n\epsilon + 3n\kappa\sqrt{\delta} \\ &= \sum_{i=1}^{n} \left[\mathbb{I}(\mathbf{X};Y_{i}|\mathbf{X}^{(\text{F})},Y^{i-1}) - \mathbb{I}(\mathbf{X};Z_{i}|\mathbf{X}^{(\text{F})},Z_{i+1}^{n}) \right] + n\epsilon + 3n\kappa\sqrt{\delta} \\ &\overset{(d)}{=} \sum_{i=1}^{n} \left[\mathbb{I}(\mathbf{X};Y_{i}|\mathbf{X}^{(\text{F})},Y^{i-1},Z_{i+1}^{n}) - \mathbb{I}(\mathbf{X};Z_{i}|\mathbf{X}^{(\text{F})},Y^{i-1},Z_{i+1}^{n}) \right] + n\epsilon + 3n\kappa\sqrt{\delta} \\ &= \sum_{i=1}^{n} \left[\mathbb{H}(Y_{i}|\mathbf{X}^{(\text{F})},Y^{i-1},Z_{i+1}^{n}) - \mathbb{H}(Y_{i}|\mathbf{X},\mathbf{X}^{(\text{F})},Y^{i-1},Z_{i+1}^{n}) - \mathbb{H}(Z_{i}|\mathbf{X}^{(\text{F})},Y^{i-1},Z_{i+1}^{n}) + \mathbb{H}(Z_{i}|\mathbf{X},\mathbf{X}^{(\text{F})},Y^{i-1},Z_{i+1}^{n}) \right] \\ &+ n\epsilon + 3n\kappa\sqrt{\delta} \\ &\overset{(e)}{=} \sum_{i=1}^{n} \left[\mathbb{H}(Y_{i}|\mathbf{X}^{(\text{F})},Y^{i-1},Z_{i+1}^{n}) - \mathbb{H}(Y_{i}|X_{i},\mathbf{X}^{(\text{F})},Y^{i-1},Z_{i+1}^{n}) - \mathbb{H}(Z_{i}|\mathbf{X}^{(\text{F})},Y^{i-1},Z_{i+1}^{n}) + \mathbb{H}(Z_{i}|X_{i},\mathbf{X}^{(\text{F})},Y^{i-1},Z_{i+1}^{n}) \right] \\ &+ n\epsilon + 3n\kappa\sqrt{\delta}. \end{split}$$

In the above, (a) and (c) follow from Lemma 4, (b) is a consequence of Fano's inequality, (d) is an application of Csiszár's sum identity [16], and (e) relies on the memoryless nature of the channel to argue that $(Y_i, Z_i) - X_i - (\mathbf{X}^{(\text{F})}, Y^{i-1}, Z^n_{i+1}, X^{i-1}, X^n_{i+1})$ is a Markov chain. Next, we let $U_i \triangleq (\mathbf{X}^{(\text{F})}, Y^{i-1}, Z^n_{i+1})$, and let T be a random variable independent of $(M, \mathbf{X}^{(\text{F})}, \mathbf{X}, \mathbf{Y}, \mathbf{Z})$ that is uniformly distributed over [1:n]. Note that $U_i - X_i - (Y_i, Z_i)$ is a Markov chain. The above inequalities are continued further as

$$nD \leq \sum_{i=1}^{n} \left[\mathbb{H}(Y_i|U_i) - \mathbb{H}(Y_i|X_i, U_i) - \mathbb{H}(Z_i|U_i) + \mathbb{H}(Z_i|X_i, U_i) \right] + n\epsilon + 3n\kappa\sqrt{\delta}$$
$$= n\mathbb{I}(X_T; Y_T|U_T, T) - n\mathbb{I}(X_T; Z_T|U_T, T) + n\epsilon + 3n\kappa\sqrt{\delta}.$$

Next, note that

$$\begin{split} \mathbb{I}(X_{T}; Z_{T}|U_{T}, T) \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_{i}; Z_{i} | \mathbf{X}^{(r)}, Y^{i-1}, Z_{i+1}^{n}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{H}(Z_{i} | \mathbf{X}^{(r)}, Y^{i-1}, Z_{i+1}^{n}) - \mathbb{H}(Z_{i} | X_{i}, \mathbf{X}^{(r)}, Y^{i-1}, Z_{i+1}^{n}) \right] \\ &\stackrel{(a)}{=} \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{H}(Z_{i} | \mathbf{X}^{(r)}, Y^{i-1}, Z_{i+1}^{n}) - \mathbb{H}(Z_{i} | \mathbf{X}, \mathbf{X}^{(r)}, Z_{i+1}^{n}) \right] \\ &\leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\mathbf{X}; Z_{i} | \mathbf{X}^{(r)}, Z_{i+1}^{n}) \\ &= \frac{1}{n} \mathbb{I}(\mathbf{X}; \mathbf{Z} | \mathbf{X}^{(r)}) \\ &\stackrel{(b)}{\leqslant} \kappa \sqrt{\delta}. \end{split}$$

In the above, (*a*) follows by noting that for each *i*, $Z_i - X_i - (X^{i-1}, X_{i+1}^n, \mathbf{X}^{(r)}, Y^{i-1}, Z_{i+1}^n)$ is a Markov chain due to the memoryless nature of the channel $P_{Z|X}$ and (*b*) follows from Lemma 4. Defining random variables (U, X, Y) with $Q_{U,X,Y}(u, x, y) = Q_{(U_T,T),X_T,Y_T}(u, x, y)$, we obtain

$$D \leq \mathbb{I}(X; Y|U) + \epsilon + 3\kappa \sqrt{\delta}.$$

Notice that $Q_{Y|X}$ is the same as the channel transition probability $P_{Y|X}$. Further, U - X - (Y,Z) is a Markov chain and $\mathbb{I}(X;Z|U) < \kappa \sqrt{\delta}$. Thus, U satisfies the constraints from the lemma statement. Finally, we bound the rate as follows.

$$nR = \mathbb{H}(M)$$

$$\stackrel{(a)}{\leq} \mathbb{I}(\mathbf{X}; \mathbf{Y}) + n\epsilon$$

$$\leq \sum_{i=1}^{n} \mathbb{I}(X_{i}; Y_{i}) + n\epsilon$$

$$= n\mathbb{I}(X_{T}; Y_{T}|T) + n\epsilon$$

$$\leq n\mathbb{I}(T, X_{T}; Y_{T}) + n\epsilon$$

$$= n\mathbb{I}(X_{T}; Y_{T}) + \mathbb{I}(T; Y_{T}|X_{T}) + n\epsilon$$

$$\stackrel{(b)}{=} n \mathbb{I}(X_T; Y_T) + n\epsilon$$
$$= n \mathbb{I}(X; Y) + n\epsilon.$$

In the above, (*a*) follows from Fano's inequality and (*b*) from the fact that $Q_{Y_T|(X_T,T)}(y|x,t) = P_{Y|X}(y|x)$. Letting $\gamma = \epsilon + 3\kappa\sqrt{\delta}$ proves the lemma.

We are now ready to formally prove Theorem 2. The converse essentially follows from the results that we have earlier in this section. Using these, we show that every achievable (R, D) must satisfy the upper bounds stated in the theorem for some choice of an auxiliary random variable U satisfying U - X - (Y,Z) and Y - U - Z. For the direct part of the proof, we prove the achievability of all (R, D) that satisfy upper bounds provided by the theorem statement when U is the zero information variable of X with respect to $P_{Z|X}$. We note that restricting the choice of U to be the zero information variable entails no loss in optimality (as shown in Lemma 3).

Proof of Theorem 2:

The converse for Theorem 2 follows by invoking Lemma 5 for a vanishing sequence of δ 's and by applying standard continuity arguments from Lemma 7 to show that any $(R, D) \in \mathscr{R}_x$ must satisfy

$$0 \le R \le \mathbb{I}(X;Y) \text{ and} \tag{16}$$

$$0 \le D \le \mathbb{I}(X; Y|U) \tag{17}$$

for some random variable U satisfying the Markov chains U - X - (Y,Z) and X - U - Z. Now, applying Lemma 3, we note that $\mathbb{I}(X;Y|U) \leq \mathbb{I}(X;Y|U_0)$, where $U_0 \in \mathscr{U}_0$ is the zero information variable of X w.r.t. $\mathsf{P}_{Z|X}$. Further, by definition, $|\mathscr{U}_0| \leq |\mathscr{X}|$. Thus, to describe the region given by Eqs. (16) and (17), it suffices to consider auxiliary variables U whose support is of size no larger than $|\mathscr{X}|$.

We now give a proof sketch for the achievability of claimed rate region. Our achievability uses a superposition code for the broadcast channel $P_{Y,Z|X}$. To this end, choose random variables (X, U) satisfying the conditions in the theorem with U as the zero information variable of X w.r.t. $P_{Z|X}$. Recall that Lemma 3 guarantees that there is no loss of optimality in choosing U as the zero information variable of X w.r.t. $P_{Z|X}$. In the following, we prove the achievability of the rate pairs that lie on the boundary of the claimed region, *i.e.*, we consider (R, D) where $R = \mathbb{I}(X; Y) - 2\epsilon$ and $D = \mathbb{I}(X; Y|U) - \epsilon$.

We consider a superposition code via a standard random coding argument. For any $\epsilon > 0$, first, we generate $\mathscr{C} = {\mathbf{u}(1), \ldots, \mathbf{u}(2^{n(R-D)})}$ by drawing $u_i(j)$ independently from the distribution P_U for each $i \in \{1, 2, \ldots, n\}$ and $j \in \{1, 2, \ldots, 2^{n(R-D)}\}$. Next, for each $j \in \{1, 2, \ldots, 2^{n(R-D)}\}$, we generate a sub-code $\mathscr{C}_j = {\mathbf{x}(j, 1), \ldots, \mathbf{x}(j, 2^{nD})}$ by drawing $x_i(j, k)$ independently from the distribution $\mathsf{P}_{X|U}(\cdot|u_i(j))$ for each $i \in \{1, 2, \ldots, n\}$ and $k \in \{1, 2, \ldots, 2^{nD}\}$. We then form the codebook $\mathscr{C} = {\mathbf{x}(\mathscr{C})(m) : m \in \mathscr{M}}$ by taking the union $\bigcup_{j \in \{1, 2, \ldots, 2^{n(R-D)}\}} \mathscr{C}_j$. Finally, the faking procedure simply accepts the transmitted codeword (say, \mathbf{x}) and outputs a uniformly drawn codeword from the sub-code that contains \mathbf{x} (say, \mathscr{C}_i).

Since the reliability of the above code follows from standard arguments for superposition coding (see [16] for example), we skip the detailed analysis here. The plausible deniability for the code follows directly from the construction by noting that for every $\mathbf{x} \in \mathcal{C}_j$, $\mathbf{u}(j)$ is precisely the sequence of the zero information symbols of \mathbf{x} w.r.t. $\mathsf{P}_{Z|X}$. Thus, for any $\mathbf{x}, \mathbf{x}' \in \mathcal{C}_j$ and $\mathbf{z} \in \mathcal{Z}^n$, $\mathsf{Q}_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{z}) = \mathsf{Q}_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}'|\mathbf{z})$.

C. Receiver Deniability

In this section, we give the proof of our achievability for Receiver Deniability in the physically degraded channel setting. As earlier, Lemma 3 shows that the rate region claimed in Theorem 3 is unchanged if *V* is restricted to be the zero information variable of *Y* with respect to $P_{YZ|X}$. In the following, we prove the achievability of (R, D) that satisfy the bounds in Theorem 3 with respect to an auxiliary variable *V* that is the zero information variable of *Y* with respect to $P_{YZ|X}$.

Proof of Theorem 3:

Let $\epsilon > 0$, fix a blocklength *n*, set

$$R = \mathbb{I}(X;Y) - \rho \tag{18}$$

for some $\rho > \epsilon$, and $|\mathcal{M}| = 2^{nR}$. Let *V* be the zero information variable of *Y* with respect to $\mathsf{P}_{YZ|X}$. Our achievability uses a random coding argument. Consider the following codebook generation procedure and the corresponding faking procedure.

a) Codebook generation: The codebook \mathscr{C} is a multiset $\{\mathbf{x}^{(\mathscr{C})}(m) : m \in \mathscr{M}\}\$ that is generated by drawing each $x_i^{(\mathscr{C})}(m)$ independently from the distribution P_X . Let $\mathsf{Pr}_{\mathscr{C}}$ be the probability distribution over the random generation of the codebook.

b) Encoding: For a message $m \in \mathcal{M}$, the encoder transmits $\mathbf{x}^{(\mathscr{C})}(m)$.

c) Decoding: Upon receiving y, the decoder looks for $m \in \mathcal{M}$ such that $(\mathbf{x}^{(\mathscr{C})}(m), \mathbf{y}) \in \mathscr{A}_{\epsilon}^{(n)}(X, Y)$.

d) Faking procedure: Given **y**, the faking procedure first generates the unique **v** where, for each *i*, v_i represents the zero information symbol of y_i w.r.t. $P_{Z|Y}$. Next, $\mathbf{Y}^{(p)}$ is drawn from \mathscr{Y}^n according to the conditional distribution $\mathbf{Q}_{\mathbf{Y}^{(p)}|\mathbf{V}} = \mathbf{Q}_{\mathbf{Y}|\mathbf{V}}$. Note that the distribution $\mathbf{Q}_{\mathbf{Y}|\mathbf{V}}$ depends on both the codebook as well the channel.

$$\begin{aligned} \mathsf{Q}_{\mathbf{Y}\mathbf{V}\mathbf{Z}}(\mathbf{y},\mathbf{v},\mathbf{z}) &\stackrel{(a)}{=} \mathsf{Q}_{\mathbf{Y}\mathbf{V}}(\mathbf{y},\mathbf{v})\mathsf{Q}_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y}) \\ &\stackrel{(b)}{=} \mathsf{Q}_{\mathbf{Y}|\mathbf{V}}(\mathbf{y}|\mathbf{v})\mathsf{Q}_{\mathbf{V}}(\mathbf{v})\mathsf{P}_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y}) \\ &\stackrel{(c)}{=} \mathsf{Q}_{\mathbf{Y}|\mathbf{V}}(\mathbf{y}|\mathbf{v})\mathsf{Q}_{\mathbf{V}}(\mathbf{v})\mathsf{P}_{\mathbf{Z}|\mathbf{V}}(\mathbf{z}|\mathbf{v}), \end{aligned}$$

and

$$\begin{split} \mathsf{Q}_{\mathbf{Y}^{(\mathrm{F})}\mathbf{VZ}}(\mathbf{y},\mathbf{v},\mathbf{z}) &= \sum_{\mathbf{y}'\in\mathscr{Y}^n} \mathsf{Q}_{\mathbf{Y},\mathbf{Y}^{(\mathrm{F})}\mathbf{VZ}}(\mathbf{y}',\mathbf{y},\mathbf{v},\mathbf{z}) \\ &\stackrel{(d)}{=} \sum_{\mathbf{y}'\in\mathscr{Y}^n} \mathsf{Q}_{\mathbf{Y}|\mathbf{V}}(\mathbf{y}'|\mathbf{v}) \mathsf{Q}_{\mathbf{Y}^{(\mathrm{F})}|\mathbf{V}}(\mathbf{y}|\mathbf{v}) \mathsf{Q}_{\mathbf{V}}(\mathbf{v}) \mathsf{Q}_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y}') \\ &\stackrel{(e)}{=} \sum_{\mathbf{y}'\in\mathscr{Y}^n} \mathsf{Q}_{\mathbf{Y}|\mathbf{V}}(\mathbf{y}'|\mathbf{v}) \mathsf{Q}_{\mathbf{Y}^{(\mathrm{F})}|\mathbf{V}}(\mathbf{y}|\mathbf{v}) \mathsf{Q}_{\mathbf{V}}(\mathbf{v}) \mathsf{P}_{\mathbf{Z}|\mathbf{V}}(\mathbf{z}|\mathbf{v}) \\ &= \mathsf{Q}_{\mathbf{Y}^{(\mathrm{F})}|\mathbf{V}}(\mathbf{y}|\mathbf{v}) \mathsf{Q}_{\mathbf{V}}(\mathbf{v}) \mathsf{P}_{\mathbf{Z}|\mathbf{V}}(\mathbf{z}|\mathbf{v}) \\ &= \mathsf{Q}_{\mathbf{Y}|\mathbf{V}|\mathbf{V}}(\mathbf{y}|\mathbf{v}) \mathsf{Q}_{\mathbf{V}}(\mathbf{v}) \mathsf{P}_{\mathbf{Z}|\mathbf{V}}(\mathbf{z}|\mathbf{v}) \\ &\stackrel{(f)}{=} \mathsf{Q}_{\mathbf{Y}|\mathbf{V}}(\mathbf{y}|\mathbf{v}) \mathsf{Q}_{\mathbf{V}}(\mathbf{v}) \mathsf{P}_{\mathbf{Z}|\mathbf{V}}(\mathbf{z}|\mathbf{v}) \\ &= \mathsf{Q}_{\mathbf{Y}\mathbf{VZ}}(\mathbf{y},\mathbf{v},\mathbf{z}). \end{split}$$

In the above, (*a*) and (*d*) follow from the dependence structure of the random variables $\mathbf{Y}, \mathbf{Y}^{(\text{F})}, \mathbf{V}$, and \mathbf{Z} , (*b*) is a consequence of the channel being physically degraded, (*c*) and (*e*) are true since *V* is the zero information variable of *Y* w.r.t. $\mathsf{P}_{Y|Z}$, and (*f*) is implied by the faking procedure used to generate $\mathbf{Y}^{(\text{F})}$. Thus,

$$\delta = \mathbb{D}(\mathsf{Q}_{\mathbf{Y}^{(\mathsf{F})}\mathbf{Z}} || \mathsf{Q}_{\mathbf{Y}\mathbf{Z}})$$

$$\leq \mathbb{D}(\mathsf{Q}_{\mathbf{Y}^{(\mathsf{F})}\mathbf{V}\mathbf{Z}} || \mathsf{Q}_{\mathbf{Y}\mathbf{V}\mathbf{Z}})$$

$$= 0.$$

Next, we analyze the rates (R, D) that our code and faking procedure can achieve. Let $\alpha \in (0, 1)$. The reliability analysis is similar to Shannon's channel coding theorem. Let $\mathcal{G}_1 \triangleq \{\mathscr{C} : \mathbb{Q}_{M,X,Y}(M \neq \hat{M}) < \epsilon\}$ denote the class of codebooks that have an average error probability smaller than ϵ . Following the standard proof of reliability of random codes, there exists $n_1 = n_1(\alpha)$ such that as long as $R < \mathbb{I}(X; Y)$ and $n > n_1$,

$$\Pr_{\mathscr{C}}(\mathcal{G}_1) \ge 1 - \alpha/4. \tag{19}$$

In the following we assume that $\mathscr{C} \in \mathcal{G}_1$ and prove that, with a high probability over the codebook generation, the rate of deniability for our faking procedure is large enough for our theorem. To this end, the following chain of inequalities give a lower bound on *D* for the code \mathscr{C} .

$$nD = \mathbb{H}(\text{DEC}(\mathbf{Y}^{(F)})|\mathbf{Y})$$

$$= \mathbb{H}(\text{DEC}(\mathbf{Y}^{(F)})|\mathbf{V}\mathbf{Y})$$

$$= \mathbb{H}(\text{DEC}(\mathbf{Y}^{(F)})|\mathbf{V})$$

$$= \mathbb{H}(\text{DEC}(\mathbf{Y})|\mathbf{V})$$

$$\geq \mathbb{I}(\text{DEC}(\mathbf{Y});\mathbf{X}|\mathbf{V})$$

$$= \mathbb{I}(\text{DEC}(\mathbf{Y}),\mathbf{Y};\mathbf{X}|\mathbf{V}) - \mathbb{I}(\mathbf{X};\mathbf{Y}|\mathbf{V},\text{DEC}(\mathbf{Y}))$$

$$= \mathbb{I}(\mathbf{X};\mathbf{Y}|\mathbf{V}) - \mathbb{I}(\mathbf{X};\mathbf{Y}|\mathbf{V},\text{DEC}(\mathbf{Y}))$$

$$\geq \mathbb{I}(\mathbf{X};\mathbf{Y}|\mathbf{V}) - n\epsilon.$$
(23)

In the above, Eq. (20) follows from the fact that **V** is a function of **Y**, (21) is due to the Markov chain $\mathbf{Y}^{(\text{F})} - \mathbf{V} - \mathbf{Y}$, and (22) follows from the faking procedure inducing $\mathbf{Q}_{\mathbf{Y}^{(\text{F})}|\mathbf{V}} = \mathbf{Q}_{\mathbf{Y}|\mathbf{V}}$. Fano's inequality implies 23 (assuming that $\mathscr{C} \in \mathcal{G}_1$). Note that the above bound is a multi-letter bound that depends on the specific codebook \mathscr{C} . A single letter bound depending only on the probability distribution of the single letter random variables follows from concentration arguments over the codebook generation process. In the following, we argue that, with high probability over the generation of \mathscr{C} , $\mathbb{I}(\mathbf{X}; \mathbf{Y}|\mathbf{V}) \ge n\mathbb{I}(X; Y|V) - n\epsilon$ for a

large enough *n*. For every $\mathbf{v} \in \mathcal{V}^n$, let us define the multi-set $\mathscr{C}_{\mathbf{v}} \triangleq \{\mathbf{x} \in \mathscr{C} : (\mathbf{x}, \mathbf{v}) \in \mathscr{A}_{\epsilon}^{(n)}(X, V)\}$.⁴ Further, for every $\mathbf{x} \in \mathscr{X}^n$, let $\mathscr{M}_{\mathbf{x}} \triangleq \{m \in \mathscr{M} : \mathbf{x}^{(\mathscr{C})}(m) = \mathbf{x}\}$. First, note that

$$\mathbb{I}(\mathbf{X}; \mathbf{Y}|\mathbf{V}) \ge \mathbb{H}(\mathbf{X}|\mathbf{V}) - n\epsilon$$

by Fano's inequality (assuming that $\mathscr{C} \in \mathcal{G}_1$). Then, given a code \mathscr{C} , there exists $\epsilon' = \epsilon'(\epsilon)$ satisfying $\lim_{\epsilon \to 0} \epsilon' = 0$ and

$$\begin{aligned}
\mathbf{H}(\mathbf{X}|\mathbf{V}) &\geq \sum_{(\mathbf{x},\mathbf{v})\in\mathscr{A}_{\epsilon}^{(n)}(X,V)} \mathbf{Q}_{\mathbf{X},\mathbf{V}}(\mathbf{x},\mathbf{v}) \log \frac{\mathbf{Q}_{\mathbf{V}}(\mathbf{v})}{\mathbf{Q}_{\mathbf{X},\mathbf{V}}(\mathbf{x},\mathbf{v})} \\
&= \sum_{(\mathbf{x},\mathbf{v})\in\mathscr{A}_{\epsilon}^{(n)}(X,V)} \mathbf{Q}_{\mathbf{X}}(\mathbf{x}) \mathbf{P}_{\mathbf{V}|\mathbf{X}}(\mathbf{v}|\mathbf{x}) \log \frac{\sum_{\mathbf{x}'\in\mathscr{C}} 2^{-nR} \mathbf{P}_{\mathbf{V}|\mathbf{X}}(\mathbf{v}|\mathbf{x}')}{\mathbf{Q}_{\mathbf{X}}(\mathbf{x}) \mathbf{P}_{\mathbf{V}|\mathbf{X}}(\mathbf{v}|\mathbf{x})} \\
&\geq \sum_{(\mathbf{x},\mathbf{v})\in\mathscr{A}_{\epsilon}^{(n)}(X,V)} \mathbf{Q}_{\mathbf{X}}(\mathbf{x}) \mathbf{P}_{\mathbf{V}|\mathbf{X}}(\mathbf{v}|\mathbf{x}) \log \frac{\sum_{\mathbf{x}'\in\mathscr{C}_{\mathbf{v}}} 2^{-nR} \mathbf{P}_{\mathbf{V}|\mathbf{X}}(\mathbf{v}|\mathbf{x}')}{\mathbf{Q}_{\mathbf{X}}(\mathbf{x}) \mathbf{P}_{\mathbf{V}|\mathbf{X}}(\mathbf{v}|\mathbf{x})} \\
&\stackrel{(a)}{\geq} \sum_{(\mathbf{x},\mathbf{v})\in\mathscr{A}_{\epsilon}^{(n)}(X,V)} \mathbf{Q}_{\mathbf{X}}(\mathbf{x}) \mathbf{P}_{\mathbf{V}|\mathbf{X}}(\mathbf{v}|\mathbf{x}) \log \frac{\sum_{\mathbf{x}'\in\mathscr{C}_{\mathbf{v}}} \mathbf{P}_{\mathbf{V}|\mathbf{X}}(\mathbf{v}|\mathbf{x}')}{|\mathscr{M}_{\mathbf{X}}| \mathbf{P}_{\mathbf{V}|\mathbf{X}}(\mathbf{v}|\mathbf{x})} \\
&\stackrel{(b)}{\geq} \sum_{(\mathbf{x},\mathbf{v})\in\mathscr{A}_{\epsilon}^{(n)}(X,V)} \mathbf{Q}_{\mathbf{X}}(\mathbf{x}) \mathbf{P}_{\mathbf{V}|\mathbf{X}}(\mathbf{v}|\mathbf{x}) \log \frac{|\mathscr{C}_{\mathbf{v}}|}{|\mathscr{M}_{\mathbf{X}}|} - 2n\epsilon'.
\end{aligned}$$
(24)

In the above, (*a*) is obtained by expressing $Q_{\mathbf{X}}(\mathbf{x})$ as $2^{-nR}|\mathscr{M}_{\mathbf{x}}|$. (*b*) follows by noting that for every (\mathbf{x}, \mathbf{v}) belonging to $\mathscr{A}_{\epsilon}^{(n)}(X, V)$, $\left|\log \frac{1}{\mathsf{P}_{V|\mathbf{X}}(\mathbf{v}|\mathbf{x})} - n\mathbb{H}(V|X)\right| < n\epsilon'$ for some $\epsilon' > 0$ that can be made arbitrarily close to 0 as ϵ approaches 0. We now show that, with high probability over the random generation of \mathscr{C} , the expression in (24) is lower bounded in the desired manner. To this end, define the following three desirable events over the codebook generation process.

$$\begin{aligned} \mathcal{G}_2 &\triangleq \left\{ \mathscr{C} : \sum_{(\mathbf{x}, \mathbf{v}) \in \mathscr{A}_{\epsilon}^{(n)}(X, V)} \mathsf{Q}_{\mathbf{X}}(\mathbf{x}) \mathsf{P}_{\mathbf{V} \mid \mathbf{X}}(\mathbf{v} \mid \mathbf{x}) > (1 - \epsilon) \right\} \\ \mathcal{G}_3 &\triangleq \left\{ \mathscr{C} : |\mathscr{C}_{\mathbf{v}}| \geq 2^{n(R - \mathbb{I}(X; V) - \epsilon'')} \ \forall \ \mathbf{v} \in \mathscr{A}_{\epsilon}^{(n)}(V) \right\} \\ \mathcal{G}_4 &\triangleq \left\{ \mathscr{C} : |\mathscr{M}_{\mathbf{x}}| < 2^{n\epsilon} \ \forall \ \mathbf{x} \in \mathscr{A}_{\epsilon}^{(n)}(X) \right\}. \end{aligned}$$

In the above, $\epsilon'' > 0$ is a constant that is specified later. Note that if $\mathscr{C} \in \bigcap_{i=1}^{4} \mathcal{G}_i$, then Eq. (23)-(24) imply that

$$D \ge \frac{1}{n} (1-\epsilon) \log \frac{2^{n(R-\mathbb{I}(X;V)-\epsilon'')}}{2^{n\epsilon}} - 2(\epsilon+\epsilon')$$

= $((1-\epsilon)(\mathbb{I}(X;Y|V) - \rho - \epsilon - \epsilon'') - 2(\epsilon+\epsilon'))$
= $\mathbb{I}(X;Y|V) - (\rho + 3\epsilon + \epsilon \mathbb{I}(X;Y|V) + 2\epsilon' + \epsilon'')$
 $\ge \mathbb{I}(X;Y|V) - (\rho + 3\epsilon + \epsilon \log |\mathscr{X}| + 2\epsilon' + \epsilon'').$ (25)

We next lower bound the probabilities of each of the above events. *i)* Event G_2 : First observe that

$$\sum_{(\mathbf{x},\mathbf{v})\in\mathscr{A}_{\epsilon}^{(n)}(X,V)} \mathsf{Q}_{\mathbf{X}}(\mathbf{x})\mathsf{P}_{\mathbf{V}|\mathbf{X}}(\mathbf{v}|\mathbf{x}) \geq \frac{|\mathscr{C}\cap\mathscr{A}_{\epsilon}^{(n)}(X)|}{|\mathscr{C}|} \min_{\mathbf{x}\in\mathscr{A}_{\epsilon}^{(n)}(X)} \sum_{\mathbf{v}\in\mathscr{A}_{\epsilon}^{(n)}(V|\mathbf{x})} \mathsf{P}_{\mathbf{V}|\mathbf{X}}(\mathbf{v}|\mathbf{x}).$$
(26)

To bound the right hand side above, we first note that using the additive form of the Chernoff bound and the definition of strong typicality,

$$\mathbb{E}_{\mathscr{C}}\left[\frac{|\mathscr{C}\cap\mathscr{A}_{\epsilon}^{(n)}(X)|}{|\mathscr{C}|}\right] = \sum_{\mathbf{x}\in\mathscr{A}_{\epsilon}^{(n)}(X)} \mathsf{P}_{\mathbf{X}}(\mathbf{x})$$

$$\geq 1 - |\mathscr{X}| \max_{\tilde{x}\in\mathscr{X}} \sum_{\mathbf{x}: \ \left|\frac{1}{n}\left|\left|\left(i:x_{i}=\tilde{x}\right)\right|-\mathsf{P}_{X}(\tilde{x})\right|\right| > \frac{\epsilon}{|\mathscr{X}|}} \mathsf{P}_{\mathbf{X}}(\mathbf{x})$$

$$\geq 1 - |\mathscr{X}| \exp\left(-\frac{n\epsilon^{2}\min_{\tilde{x}\in\mathscr{X}}\mathsf{P}_{X}(\tilde{x})}{4|\mathscr{X}|^{2}}\right). \tag{27}$$

⁴Recall that \mathscr{C} is a multi-set with possibly repeated elements. As a result, \mathscr{C}_{v} may also contain codewords that have multiplicity greater than one.

In particular, for a large enough n, we have

$$\mathbb{E}_{\mathscr{C}}\left[\frac{|\mathscr{C} \cap \mathscr{A}_{\epsilon}^{(n)}(X)|}{|\mathscr{C}|}\right] \ge 1 - \epsilon/4.$$

Next, by standard properties of the conditionally typical set, we have, for large enough n,

$$\sum_{\boldsymbol{\epsilon} \in \mathscr{A}_{\epsilon}^{(n)}(\boldsymbol{V}|\mathbf{x})} \mathsf{P}_{\mathbf{V}|\mathbf{X}}(\mathbf{v}|\mathbf{x}) \ge 1 - \epsilon/4.$$
(28)

Combining (27) and (28), we conclude that there exists n^* such that for every $n > n^*$,

$$\mathbb{E}_{\mathscr{C}}\left[\frac{|\mathscr{C}\cap\mathscr{A}_{\epsilon}^{(n)}(X)|}{|\mathscr{C}|}\min_{\mathbf{x}\in\mathscr{A}_{\epsilon}^{(n)}(X)}\sum_{\mathbf{v}\in\mathscr{A}_{\epsilon}^{(n)}(V|\mathbf{x})}\mathsf{P}_{\mathbf{V}|\mathbf{X}}(\mathbf{v}|\mathbf{x})\right] > 1 - \frac{\epsilon}{2}.$$

The above expression gives, in expectation, a lower bound on the left hand side of (26). A further concentration argument over the i.i.d. generation of the codebook shows the existence of $n_2 = n_2(\epsilon)$ such that whenever $n > n_2(\alpha)$,

$$\Pr_{\mathscr{C}}(\mathcal{G}_2) \ge 1 - \alpha/4. \tag{29}$$

ii) Event \mathcal{G}_3 : Next, note that for any $\mathbf{v} \in \mathscr{A}_{\epsilon}^{(n)}(V)$, there exists $n^{\#}$ and $\epsilon'' = \epsilon''(\epsilon)$ satisfying $\lim_{\epsilon \to 0} \epsilon'' = 0$ and

$$\mathbb{E} |\mathscr{C}_{\mathbf{v}}| = 2^{nR} \sum_{\mathbf{x}: (\mathbf{x}, \mathbf{v}) \in \mathscr{A}_{\epsilon}^{(n)}(X, V)} \mathsf{P}_{\mathbf{X}}(\mathbf{x})$$
$$\geq 2^{n(R - \mathbb{I}(X; V) - \epsilon''/2)}$$

whenever $n > n^{\#}$. Now, since each codeword falls in \mathcal{C}_{v} in an independent and identical manner over the codebook generation, the true value of \mathcal{C}_{v} concentrates around its mean with a high probability. In particular, by applying Chernoff bound on \mathcal{C}_{v} , we obtain that there exists $n_{3} = n_{3}(\epsilon)$ such that for every $n > n_{3}(\alpha)$,

$$\Pr_{\mathscr{C}}(\mathcal{G}_{3}) \ge \Pr_{\mathscr{C}}(|\mathscr{C}_{\mathbf{v}}| \ge 2^{-n\epsilon^{n}/2}\mathbb{E}|\mathscr{C}_{\mathbf{v}}|)$$

> 1 - \alpha/4. (30)

iii) Event \mathcal{G}_4 : Finally, let $\beta = 2^{n\epsilon}$, and observe that there exists $\epsilon''' = \epsilon'''(\epsilon)$ such that $\lim_{\epsilon \to 0} \epsilon'''(\epsilon) = 0$ and

$$\log (\Pr_{\mathscr{C}}(\mathscr{C} \notin \mathcal{G}_{4})) = \log \Pr_{\mathscr{C}}(\exists \mathscr{S} \subseteq \mathscr{M}, \mathbf{x} \in \mathscr{A}_{\epsilon}^{(m)}(X) \text{ s.t. } |\mathscr{S}| = \beta \text{ and } \mathbf{x}^{(\mathfrak{C})}(m) = \mathbf{x} \forall m \in \mathscr{S})$$

$$\leq \log \sum_{\substack{\mathscr{S} \subseteq \mathscr{M} \\ |\mathscr{S}| = \beta}} \sum_{\mathbf{x} \in \mathscr{A}_{\epsilon}^{(m)}(X)} \prod_{m \in \mathscr{S}} \Pr_{\mathscr{C}}\left(\mathbf{x}^{(\mathscr{C})}(m) = \mathbf{x}\right)$$

$$= \log \left(\frac{|\mathscr{M}|}{\beta} \right) + \log \sum_{\mathbf{x} \in \mathscr{A}_{\epsilon}^{(m)}(X)} \left(\Pr_{\mathscr{C}}\left(\mathbf{x}^{(\mathscr{C})}(1) = \mathbf{x}\right)\right)^{\beta}$$

$$\stackrel{(a)}{\leq} |\mathscr{M}|H_{b}\left(\frac{\beta}{|\mathscr{M}|}\right) + \log |\mathscr{A}_{\epsilon}^{(n)}(X)| + \beta \log \max_{\mathbf{x} \in \mathscr{A}_{\epsilon}^{(m)}(X)} \Pr_{\mathscr{C}}(\mathbf{x}^{(\mathscr{C})}(1) = \mathbf{x})$$

$$\stackrel{(b)}{\leq} \beta \log \frac{|\mathscr{M}|}{\beta} + (|\mathscr{M}| - \beta) \log \frac{|\mathscr{M}|}{|\mathscr{M}| - \beta} - (\beta - 1)n\mathbb{H}(X) + (\beta + 1)n\epsilon'''$$

$$\stackrel{(c)}{\leq} \beta \log \frac{|\mathscr{M}|}{\beta} + \beta \log e - (\beta - 1)n\mathbb{H}(X) + (\beta + 1)n\epsilon'''$$

$$= 2^{n\epsilon}(n(R - \epsilon) + \log e) - (2^{n\epsilon} - 1)n\mathbb{H}(X) + (2^{n\epsilon+1})n\epsilon'''$$

$$= 2^{n\epsilon}(n(R - \mathbb{H}(X) - \epsilon + \epsilon''') + \log e) + n(\mathbb{H}(X) + \epsilon''')$$

$$\leq 2^{n\epsilon}(n(R - \mathbb{I}(X; Y) - \epsilon + \epsilon''') + \log e) + n(\mathbb{H}(X) + \epsilon''')$$

$$= 2^{n\epsilon}(n(-\rho - \epsilon + \epsilon''') + \log e) + n(\mathbb{H}(X) + \epsilon'''). \tag{31}$$

In the above, (*a*) is a standard upper bound on $\binom{|\mathcal{M}|}{\beta}$ in terms of the binary entropy function $H_b(\beta/|\mathcal{M}|)$. (*b*) obtained by noting that there exists ϵ''' such that $\lim_{\epsilon \to 0} \epsilon''' = 0$, $|\mathscr{A}_{\epsilon}^{(n)}(X)| \leq 2^{n(\mathbb{H}(X) + \epsilon''')}$ and $\mathsf{P}_{\mathbf{X}}(\mathbf{x}) \leq 2^{-n(\mathbb{H}(X) - \epsilon''')}$ for each $\mathbf{x} \in \mathscr{A}_{\epsilon}^{(n)}(X)$. Lastly, (*c*) is obtained by using the fact that for every a > 0, $\log a = \log e \ln a \leq (a - 1) \log e$. Note that as long as ρ is strictly greater than $\epsilon''' - \epsilon$, the right hand side of (31) diverges to $-\infty$ as *n* increases without bound. In particular, this implies that there exists n_4 such that for every $n > n_4(\alpha)$,

$$\Pr_{\mathscr{C}}(\mathcal{G}_4) > 1 - \alpha/4. \tag{32}$$

Finally, combining (19), (29), (30), and (32) we conclude that, whenever $n > \max\{n_1, n_2, n_3, n_4\}$, with probability at least $1 - \alpha$, the randomly drawn code is simultaneously (ϵ , R)-reliable and (0, D)-plausibly deniable where (R, D) satisfy the lower bounds in (18) and (25). Since ρ and ϵ can be made arbitrarily close to zero, this shows the achievability of all rates in the interior of the claimed region.

D. Discussions

1) An example:

Example 3. Consider a channel $P_{YZ|X}$ with $\mathscr{X} = \mathscr{Y} = \mathscr{X} = \{1, 2, 3\}, Y = X$ and $P_{Z|X}$ as in Figure 6, *i.e.*,

$$\mathsf{P}_{YZ|X}(y,z|x) = \begin{cases} 0.3 & (x,y,z) \in \{(1,1,1),(2,2,1)\} \\ 0.7 & (x,y,z) \in \{(1,1,2),(2,2,2)\} \\ 0.4 & (x,y,z) = (3,3,2) \\ 0.6 & (x,y,z) = (3,3,3) \\ 0 & \text{otherwise.} \end{cases}$$

We characterize the capacity region $\mathscr{R}_{\mathbf{x}}$ by restricting our choice of the auxililary random variable U to the zero-information random variable. For the above conditional distribution, the zero-information random variable of X w.r.t. $\mathsf{P}_{Z|X}$ takes two values: $u_1 = \{1, 2\}$ and $u_2 = \{3\}$. Since X = Y, $\mathbb{I}(X; Y) = \mathbb{H}(X)$ and $\mathbb{I}(X; Y|U) = \mathsf{P}_X(1)\log\frac{\mathsf{P}_X([1,2])}{\mathsf{P}_X(1)} - \mathsf{P}_X(2)\log\frac{\mathsf{P}_X([1,2])}{\mathsf{P}_X(2)}$. The capacity region $\mathscr{R}_{\mathbf{x}}$ (Figure 7) consists of all (R, D) pairs satisfying the following

$$D \le R \le H_t\left(\frac{D}{2}, \frac{D}{2}, 1-D\right)$$
$$0 \le D \le 1,$$

where $H_t(\cdot, \cdot, \cdot)$ represents the ternary entropy function. Interestingly, the capacity region depends on the conditional distribution $P_{Z|X}$, only through the zero-information variable induced by it – all conditional distributions $P_{Z|X}$ that induce the same zero-information variable have the same capacity region (assuming $P_{Y|X}$ is unchanged). This is a general feature of capacity regions for the transmitter deniability problem.



Fig. 7: Capacity region $\mathscr{R}_{\mathbf{x}}$ for Example 3.

2) Rate of deniability as the Equivocation rate: Similar to the Message Deniability setting, we can attach a secrecy interpretation to the rate of deniability for faking procedures that are plausibly deniable. The following proposition mirrors Proposition 2.

Proposition 3. Let $\mathbf{X}^{(\text{F})}$ be (δ, D) -plausibly deniable for \mathbf{X} given \mathbf{Z} and satisfy the Markov chain $\mathbf{X}^{(\text{F})} - \mathbf{X} - \mathbf{Z}$. Then, there exists $\mu \ge 0$ depending only on $\mathsf{P}_{Z|X}$ such that

$$nD - n\mu\sqrt{\delta} \leq \mathbb{H}(M|\mathbf{Z}, \mathbf{X}^{(\mathrm{F})}) \leq nD + n\mu\sqrt{\delta}.$$

Proof:

The proof relies on the Lemma 4 and proceeds in similar spirit as Proposition 2. To this end, let κ be the constant defined in Lemma 4. Note that

$$\mathbb{H}(M|\mathbf{Z}, \mathbf{X}^{(\mathrm{F})}) = \mathbb{H}(M|\mathbf{X}^{(\mathrm{F})}) - \mathbb{I}(M; \mathbf{Z}|\mathbf{X}^{(\mathrm{F})})$$

and

$$\mathbb{H}(M|\mathbf{Z},\mathbf{X}^{(\mathrm{F})}) \geq nD - 2n\kappa\sqrt{\delta}.$$

Choosing $\mu = 2\kappa$ completes the proof.

VI. CONCLUDING REMARKS

In this paper, we have considered three different models of Plausible Deniability and give achievable rates for each model while also giving tight converses for the message deniability and transmitter deniability settings. It is evident that, at the very least, each capacity region is a subset of the *Rate-Equivocation* region. Intuitively, this may be interpreted as follows – any code that has a rate of deniability D has the property that the equivocation at the eavesdropper is at least D (otherwise, with high probability, the eavesdropper can detect a fake response). On the other hand, it is not *a priori* clear whether the achievable rates for any one model considered in this paper is a subset of another – part of the difficulty in comparing the different settings arises from the fact that in each setting, the faking procedure accepts different inputs to generate the fake output.

Digging deeper into the nature of our problem, our achievability proofs rely crucially on the summoned party's ability to identify a set of *plausible* fake responses that appear roughly as likely as the true response to an eavesdropper who also observes the channel output. Further, the set of plausible responses must be identified without knowing the eavesdropper actual channel observation. To achieve this goal, our schemes ensure that the set of possible response values partitions into "cliques" such that each response from the clique would be plausible to the eavesdropper given any likely channel output. This simplifies our faking procedure to randomly picking one response from the clique corresponding to the true response. In our scheme for the message deniability setting, these cliques correspond to all messages that are consistent with the transmitted "public message", while in the transmitter and receiver deniability settings, these cliques correspond to codewords and received vectors that are statistically consistent with the zero information variables of the actual transmitted codeword and the received vector, respectively. In each of these settings, given the clique corresponding to the true value of the summoned party's response, the eavesdropper's channel observation provides asymptotically negligible additional information about the true value of the response.

Given our problem formulation, the above achievability idea appears natural. Perhaps surprisingly, we also show that any good faking procedure for our problem must follow the above decomposition (at least roughly). In the transmitter and receiver deniability settings, Lemmas 2 and 4 make this claim precise. A drastic consequence of this is that non-zero rates are possible for transmitter deniability only when non-trivial zero information variables exist with respect to the eavesdropper's channel output. We note that the existence of such variables is guaranteed only for fairly special classes of channels – even for channels such as Binary Symmetric Channels, the only zero information variables are the channel inputs themselves. Further, the existence of non-trivial zero information variables may be rather fragile with respect to perturbations in the channel conditional probability. This is in contrast to the message deniability setting, the capacity region for plausible deniability seems somewhat robust to the channel statistics (*c.f.* [18] for the robustness analysis for a related problem).

Our work potentially leads to several intriguing open questions. In settings where non-zero rates of deniability are not possible (*e.g.* transmitter deniability over a binary symmetric broadcast channel), it is of interest to understand whether an asymptotically vanishing rate of communication may still be possible. Recent work on "*square-root law*" in covert communications [11]–[14] suggests such a possibility. However, unlike covert communication, the eavesdropper in our setting has potentially greater distinguishing power due to access to both the channel observation and the summoned party's response. Separately, while our work examines the broadcast channel setting, the notion of information theoretic plausible deniability readily extends to other communication settings with security oriented goals, *e.g.*, secret key generation, interactive communication, and communication with public discussion. It would be interesting to examine the capacity question in these settings. Finally, we remark that while our formulation of plausible deniability relies on the asymmetry between the channel to the eavesdropper and the legitimate receiver, and the cryptographic formulation of [6], [8], [10] relies on the eavesdropper's inability to efficiently compute certain functions without knowing the receiver's private key, it would be interesting to obtain plausibly deniable communication.

APPENDIX A

STRONG SECRECY FOR BROADCAST CHANNELS WITH BOTH CONFIDENTIAL AND LEAKED MESSAGES

In the following, we consider the problem of broadcast channel with both confidential and leaked messages described in Figure 3. We first give the proof of Lemma 1 that gives an inner bound on the capacity region defined in Definition 1.

Proof of Lemma 1:

The proof essentially follows from the strategies used in [19, Theorem 17.13], [20, Theorem 3], [21] to prove strong secrecy capacity region for the setting of *broadcast channel with confidential messages* (Figure 4). The only difference here from the settings of [19]–[21], is that we do not demand that the message *t* be reliably decoded by Judy from her observation \mathbf{z} . This allows us to send *t* at all rates less than $\mathbb{I}(V; Y)$ instead of min { $\mathbb{I}(V; Y)$, $\mathbb{I}(V; Z)$ } for every (*U*, *V*) pair satisfying the lemma conditions. As the proof would be nearly identical to the proofs supplied in [19]–[21], we skip the full proof of the lemma here.

Next, we give a lemma that allows us to modify the strong secrecy metric (condition 3 of Definition 1) to the Kullback-Leibler Divergence in the form suitable for our problem.

Lemma 6. Let $\alpha \in (0, 1)$ and $\beta > 0$. Let \tilde{I}, J be random variables distributed on I and \mathcal{J} respectively with a joint distribution $\mathsf{P}_{\tilde{I},J}$ such that $\mathbb{I}(\tilde{I};J) < \beta$. Then, there exists a random variable $I \in I$ that is jointly distributed with \tilde{I} and J in accordance with a Markov chain $I - \tilde{I} - J$ such that the joint distribution $\mathsf{P}_{I,\tilde{I},J}$ has the following properties:

1) $\mathsf{P}_{I,\tilde{I}}(I \neq \tilde{I}) > 1 - \alpha$. 2) $\mathsf{P}_{I}(i) = \mathsf{P}_{\tilde{I}}(i)$ for all $i \in I$.

$$\frac{2}{3} I(I:I) < \beta$$

4) $\mathbb{D}(\mathsf{P}_I \mathsf{P}_I || \mathsf{P}_{I,I}) \le \sqrt{2\beta} \log(1/\alpha).$

Proof:

In the following, we assume, without loss of generality, that $\mathsf{P}_{\tilde{I}}(\tilde{i}) > 0$ and $\mathsf{P}_{J}(j) > 0$ for each $(\tilde{i}, j) \in I \times \mathcal{J}$. We construct the random variable *I* explicitly as follows. First, we define the transition probability

$$\mathsf{P}_{I|\tilde{I}}(i|\tilde{i}) = \begin{cases} 1 - \alpha + \alpha \mathsf{P}_{\tilde{I}}(\tilde{i}) & i = \tilde{i} \\ \alpha \mathsf{P}_{\tilde{I}}(i) & i \neq \tilde{i}, \end{cases}$$

and let $\mathsf{P}_{I,\tilde{I},J}(i,\tilde{i},j) = \mathsf{P}_{I|\tilde{I}}(i|\tilde{i})\mathsf{P}_{\tilde{I},J}(\tilde{i},j)$ for all $(i,\tilde{i},j) \in I \times I \times J$. Clearly, I equals \tilde{I} with probability at least $1 - \alpha$. Hence, condition 1 is satisfied. Also, $\mathsf{P}_{I}(i) = \sum_{\tilde{i} \in I} \mathsf{P}_{I|\tilde{I}}(i|\tilde{i})\mathsf{P}_{\tilde{I}}(\tilde{i}) = \mathsf{P}_{\tilde{I}}(i)$, which implies that condition 2 is also satisfied. Further, noting that $I - \tilde{I} - J$ is a Markov chain, by the Data Processing inequality,

$$\mathbb{I}(I;J) \le \mathbb{I}(\tilde{I};J) < \beta.$$

Thus, condition 3 is satisfied as well. Note that

$$\mathsf{P}_{I,J}(i,j) = \sum_{\tilde{i}\in I} \mathsf{P}_{I|\tilde{i}}(i|\tilde{i})\mathsf{P}_{\tilde{I},J}(\tilde{i},j)$$

= $(1 - \alpha + \alpha\mathsf{P}_{I}(i))\mathsf{P}_{\tilde{I},J}(i,j) + \alpha\mathsf{P}_{I}(i)\sum_{\tilde{i}\in I\setminus\{i\}}\mathsf{P}_{\tilde{I},J}(\tilde{i},j)$
 $\geq \alpha\mathsf{P}_{I}(i)\mathsf{P}_{I}(j).$ (33)

Note that, by our assumption, $P_I(i)P_J(j) > 0$ for each $(i, j) \in I \times \mathcal{J}$. Further, Eq. (33) implies that $P_{I,J}(i, j) > 0$ for each $(i, j) \in I \times \mathcal{J}$. Thus, $\mathbb{D}(\mathsf{P}_I \mathsf{P}_J || \mathsf{P}_{I,J})$ and $\mathbb{D}(\mathsf{P}_{I,J} || \mathsf{P}_I \mathsf{P}_J)$ are finite and well-defined. Now,

$$\begin{split} \mathbb{D}(\mathsf{P}_{I}\mathsf{P}_{J}||\mathsf{P}_{I,J}) &= \sum_{i\in\mathcal{I},j\in\mathcal{J}}\mathsf{P}_{I}(i)\mathsf{P}_{J}(j)\log\frac{\mathsf{P}_{I}(i)\mathsf{P}_{J}(j)}{\mathsf{P}_{I,J}(i,j)} \\ &= \sum_{i\in\mathcal{I},j\in\mathcal{J}}(\mathsf{P}_{I}(i)\mathsf{P}_{J}(j) - \mathsf{P}_{I,J}(i,j))\log\frac{\mathsf{P}_{I}(i)\mathsf{P}_{J}(j)}{\mathsf{P}_{I,J}(i,j)} + \sum_{i\in\mathcal{I},j\in\mathcal{J}}\mathsf{P}_{I,J}(i,j)\log\frac{\mathsf{P}_{I}(i)\mathsf{P}_{J}(j)}{\mathsf{P}_{I,J}(i,j)} \\ &\stackrel{(a)}{\leq} ||\mathsf{P}_{I,J} - \mathsf{P}_{I}\mathsf{P}_{J}||_{1}\max_{i\in\mathcal{I},j\in\mathcal{J}}\log\frac{\mathsf{P}_{I}(i)\mathsf{P}_{J}(j)}{\mathsf{P}_{I,J}(i,j)} - \mathbb{D}(\mathsf{P}_{IJ}||\mathsf{P}_{I}\mathsf{P}_{J}) \\ &\stackrel{(b)}{\leq}\sqrt{2\beta}\log(1/\alpha). \end{split}$$

In the above, (a) follows from Hölder's inequality. (b) is obtained by applying the non-negativity of the Kullback-Leibler Divergence, inequality (33), and by noting that

$$\begin{split} \|\mathsf{P}_{I,J} - \mathsf{P}_{I}\mathsf{P}_{J}\|_{1} &= \sum_{i \in \mathcal{I}, j \in \mathcal{J}} \left| \sum_{\tilde{i} \in \mathcal{I}} \mathsf{P}_{I|\tilde{I}}(i|\tilde{i}) \left(\mathsf{P}_{\tilde{I},J}(\tilde{i},j) - \mathsf{P}_{\tilde{I}}(\tilde{i})\mathsf{P}_{J}(j) \right) \right| \\ &\leq \sum_{i \in \mathcal{I}, \tilde{i} \in \mathcal{I}, j \in \mathcal{J}} \mathsf{P}_{I|\tilde{I}}(i|\tilde{i}) \left| \mathsf{P}_{\tilde{I},J}(\tilde{i},j) - \mathsf{P}_{\tilde{I}}(\tilde{i})\mathsf{P}_{J}(j) \right| \\ &= \|\mathsf{P}_{\tilde{I},J} - \mathsf{P}_{\tilde{I}}\mathsf{P}_{J}\|_{1} \end{split}$$

$$\stackrel{(a)}{\leq} \sqrt{2\mathbb{D}\left(\mathsf{P}_{\tilde{I},J} \| \mathsf{P}_{\tilde{I}}\mathsf{P}_{J}\right)} \\ \leq \sqrt{2\beta}.$$

In the above, (a) follows from Pinsker's inequality. This proves that I satisfies the conditions 2-4.

Finally, we give a proof of Corollary 1.

Proof of Corollary 1:

The proof follows by starting with a code from Lemma 1 and using Lemma 6 to modify it to achieve the desired properties. Let $\tilde{\mathscr{C}}$ be a code that satisfies conditions 1-3 of Definition 1. Let $\tilde{S} \in \mathscr{S}$ denote the confidential message and $T \in \mathscr{T}$ denote the leaked message for this code. Note that the random variables \tilde{S}, T , and \mathbb{Z} satisfy $\mathbb{I}(\tilde{S}; T, \mathbb{Z}) < \delta$. Next, apply Lemma 6 with \tilde{S} , (T, \mathbb{Z}) , ϵ , and δ , in place of \tilde{I} , J, α , and β , respectively to obtain the random variable S (in place of I) that is jointly distributed with \tilde{S} and (T, \mathbb{Z}) according to a distribution $\mathbb{Q}_{S,\tilde{S},(T,\mathbb{Z})} = \mathbb{Q}_{S|\tilde{S}} \mathbb{Q}_{\tilde{S}} \mathbb{Q}_{(T,\mathbb{Z})|\tilde{S}}$.

Consider a code \mathscr{C} that operates as follows. Let (S, T) be the messages for this code. First, Alice maps the message S to a randomly drawn \tilde{S} according to the transition probability $Q_{\tilde{S}|S}$. Next, she encodes (\tilde{S}, T) using the encoder for \mathscr{C} . Upon receiving **Y**, Bob uses the decoder for \mathscr{C} to output his reconstruction of (S, T).

By Lemma 6, the overall code satisfies the conditions of Definition 1 with requirement 2 replaced by

(y.

$$\sum_{(s,t): \text{DEC}(\mathbf{y}) \neq (s,t)} \mathsf{Q}_{\mathbf{Y},S,T}(\mathbf{y},s,t) \le 2\epsilon$$

In addition, the code also satisfies the following property

$$\mathbb{D}(\mathsf{Q}_{S}\mathsf{Q}_{T,\mathbf{Z}}||\mathsf{Q}_{S,T,\mathbf{Z}}) < \sqrt{2\delta}\log(1/\epsilon).$$

Now, by first choosing ϵ small enough and subsequently, δ small enough, both the error probability and the K-L divergence above can be made arbitrarily small. This proves the corollary.

APPENDIX B

A CONTINUITY PROPERTY

Lemma 7. Let \mathscr{P} be a compact subset of the set of probability measures over a finite set \mathscr{B} . Let $\mathbb{L} : \mathscr{P} \to \mathbb{R}^+$ and $\mathbb{M} : \mathscr{P} \to \mathbb{R}^+$ be functionals that are continuous with respect to the variational distance such that $\mathbb{L}^{-1}(\{0\}) \neq \phi$. Then,

$$\lim_{\delta \to 0^+} \max_{\mathsf{P} \in \mathscr{P}: \mathbb{L}(\mathsf{P}) < \delta} \mathbb{M}(\mathsf{P}) = \max_{\mathsf{P} \in \mathscr{P}: \mathbb{L}(\mathsf{P}) = 0} \mathbb{M}(\mathsf{P}).$$
(34)

Proof:

Since \mathscr{P} is compact and \mathbb{M} is a continuous on \mathscr{P} , \mathbb{M} is bounded. Further, as $\mathbb{M}(\mathsf{P}) \ge 0$ for every $\mathsf{P} \in \mathscr{P}$, and $\max_{\mathsf{P} \in \mathscr{P}: \mathbb{L}(\mathsf{P}) < \delta} \mathbb{M}(\mathsf{P})$ is an increasing function of δ , the limit on the left hand side of Eq (34) exists. Now, for any $\delta > 0$,

$$\max_{\mathsf{P}\in\mathscr{P}:\mathbb{L}(\mathsf{P})<\delta}\mathbb{M}(\mathsf{P})\geq \max_{\mathsf{P}\in\mathscr{P}:\mathbb{L}(\mathsf{P})=0}\mathbb{M}(\mathsf{P})$$

Taking the limit as δ approaches zero, the left hand side of Eq. (34) is at least as large as the right hand side. Next, we show that the limit on the left hand side cannot be larger than the right hand side.

To this end, let $M^* = \lim_{\delta \to 0} \max_{\mathsf{P} \in \mathscr{P}: \mathbb{L}(\mathsf{P}) < \delta} \mathbb{M}(\mathsf{P})$. Thus, there exists a sequence $\{\mathsf{P}^{(i)}\}_{i \in \mathbb{N}}$ in \mathscr{P} such that $\mathbb{L}(\mathsf{P}^{(i)}) < 1/i$ and $\lim_{i \to \infty} \mathbb{M}(\mathsf{P}^{(i)}) = M^*$. As \mathscr{P} is a compact set under the variational distance, $\{\mathsf{P}^{(i)}\}_{i \in \mathbb{N}}$ contains a subsequence $\{\mathsf{P}^{(i_j)}\}_{j \in \mathbb{N}}$ that converges (in variational distance) to a limiting distribution P^* . By continuity of \mathbb{L} , we have

$$0 \leq \mathbb{L}(\mathsf{P}^*_{\mathbf{B}}) = \lim_{i \to \infty} \mathbb{L}(\mathsf{P}^{(i_j)}) \leq \lim_{i \to \infty} 1/i_j = 0.$$

Thus, $M^* = \mathbb{M}(\mathsf{P}^*) \le \max_{\{\mathsf{P} \in \mathscr{P} : \mathbb{L}(\mathsf{P}) = 0\}} \mathbb{M}(\mathsf{P}).$

REFERENCES

- [1] J. Katz and Y. Lindell, Introduction to Modern Cryptography. Chapman & Hall/CRC, 2007.
- [2] A. Wyner, "The Wire-tap Channel," Bell System Technical Journal, The, vol. 54, no. 8, pp. 1355–1387, Oct 1975.
- [3] I. Csiszar and J. Körner, "Broadcast Channels with Confidential Messages," *IEEE Transactions on Information Theory*, vol. 24, no. 3, pp. 339–348, May 1978.
- [4] M. Bloch and J. Barros, Physical-Layer Security: From Information Theory to Security Engineering. Cambridge University Press, 2011.
- [5] N. Cai and R. W. Yeung, "Secure network coding," in *Information Theory*, 2002. Proceedings. 2002 IEEE International Symposium on, 2002, pp. 323–.
 [6] R. Canetti, C. Dwork, M. Naor, and R. Ostrovsky, "Deniable Encryption," in *Proceedings of the 17th Annual International Cryptology Conference on Advances in Cryptology*. London, UK: Springer-Verlag, 1997, pp. 90–104.
- [7] J. Benaloh and D. Tuinstra, "Uncoercible communication," Computer Science Technical Report TR-MCS-94-1, Clarkson University, 1994.
- [8] A. O'Neill, C. Peikert, and B. Waters, *Bi-Deniable Public-Key Encryption*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 525–542. http://dx.doi.org/10.1007/978-3-642-22792-9_30
- [9] "Truecrypt: Hidden operating system," http://truecrypt.sourceforge.net.

- [10] A. Sahai and B. Waters, "How to Use Indistinguishability Obfuscation: Deniable Encryption, and More," in Proceedings of the 46th Annual ACM Symposium on Theory of Computing. New York, NY, USA: ACM, 2014, pp. 475-484.
- [11] B. Bash, D. Goeckel, and D. Towsley, "Limits of Reliable Communication with Low Probability of Detection on AWGN Channels," IEEE Journal on Selected Areas in Communications, vol. 31, no. 9, pp. 1921–1930, September 2013.
- [12] P. H. Che, M. Bakshi, and S. Jaggi, "Reliable Deniable Communication: Hiding Messages in Noise," in Proceedings of the 2013 IEEE International Symposium on Information Theory, July 2013, pp. 2945–2949.
- [13] M. R. Bloch, "Covert communication over noisy channels: A resolvability perspective," IEEE Transactions on Information Theory, vol. 62, no. 5, pp. 2334-2354, May 2016.
- [14] L. Wang, G. W. Wornell, and L. Zheng, "Fundamental limits of communication with low probability of detection," IEEE Transactions on Information Theory, vol. 62, no. 6, pp. 3493-3503, June 2016.
- [15] P. H. Che, M. Bakshi, C. Chan, and S. Jaggi, "Reliable, deniable and hidable communication," in 2014 Information Theory and Applications Workshop (ITA), Feb 2014, pp. 1-10.
- [16] A. El Gamal and Y.-H. Kim, Network Information Theory. Cambridge University Press, 2011.
 [17] R. Ahlswede and J. Korner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Transactions on Information* Theory, vol. 21, no. 6, pp. 629-637, November 1975.
- [18] R. F. Schaefer and H. Boche, "Robust broadcasting of common and confidential messages over compound channels: Strong secrecy and decoding performance," IEEE Transactions on Information Forensics and Security, vol. 9, no. 10, pp. 1720–1732, Oct 2014.
- [19] I. Csiszar and J. Körner, Information Theory: Coding Theorems for Discrete Memoryless Systems. Cambridge University Press, 2011.
- [20] M. R. Bloch and J. N. Laneman, "Strong secrecy from channel resolvability," IEEE Transactions on Information Theory, vol. 59, no. 12, pp. 8077–8098, Dec 2013.
- [21] R. Matsumoto and M. Hayashi, "Strong security and separated code constructions for the broadcast channels with confidential messages," CoRR, vol. abs/1010.0743, 2010. http://arxiv.org/abs/1010.0743