Reliable Deniable Communication: Hiding Messages in Noise

Pak Hou Che, Mayank Bakshi, Sidharth Jaggi
The Chinese University of Hong Kong

***Abstract*—Alice may wish to *reliably* send a message to Bob over a binary symmetric channel (BSC) while ensuring that her transmission is *deniable* from an eavesdropper Willie. That is, if Willie observes a “significantly noisier” transmission than Bob does, he should be unable to estimate even whether Alice is transmitting or not. Even when Alice’s (potential) communication scheme is publicly known to Willie (with *no* common randomness between Alice and Bob), we prove that over n channel uses Alice can transmit a message of length $\mathcal{O}(\sqrt{n})$ bits to Bob, deniably from Willie. We also prove information-theoretically order-optimality of our results.**

I. Introduction

Alice is in jail, and may wish to communicate reliably with Bob in the neighboring cell, over n uses of a noisy BSC (if she stays silent, the input to the channel is all zeroes). Unfortunately, the warden Willie is monitoring Alice (though his observations are significantly noisier, since his CCTV camera is low-quality).[1] Willie only wishes to detect Alice’s “transmission status” (*i.e.*, he only wants to know whether she’s talking or not, and doesn’t necessarily care *what* she’s saying). Hence Alice wishes to use a communication scheme that is “deniable from Willie”, *i.e.* Willie’s *best* estimate of Alice’s transmission status should be essentially statistically independent of his observations.In this work we show:

1. *Deniability – outer bound on codeword weight (Thm 1)*: If the binary code Alice uses to encode her message contains a substantial fraction of “high-weight codewords” (that is, have weight that is $\omega(\sqrt{n})$ over n channel uses), then her communication scheme cannot be deniable. In particular, Willie can simply count the number of non-zero symbols he observes, and compare this number with a simple function of channel parameters, to estimate, fairly accurately, Alice’s transmission status. Hence, for deniability from Willie, Alice’s code should comprise mostly of “low-weight codewords”.

2. *Reliability/deniability – throughput outer bound (Thm 2)*: Since the link to Bob is *also* noisy, Alice needs to code. We use information-theoretic inequalities to demonstrate that for any code that is simultaneously highly deniable and reliable, a message of at most $\mathcal{O}(\sqrt{n})$ bits can be transmitted by Alice over n channel uses. [2]

3. *Reliability/deniability – achievable scheme (Thm 3)*: If Willie’s BSC is “sufficiently noisier” (in a precise sense that we quantify later) than Bob’s BSC, then we design a communication scheme (publicly known to all parties – Alice, Bob and Willie) such that:

- *Throughput*: It encodes a message with $\mathcal{O}(\sqrt{n})$ bits in n channel uses (or, when the channel to Bob is noiseless, though the channel to Willie is still noisy, $\mathcal{O}(\sqrt{n}\log n)$ bits).
- *Reliability*: It enables Bob to correctly reconstruct Alice’s message with high probability.
- *Deniability*: Willie’s best estimate of Alice’s transmission status is essentially statistically independent of his observations of his channel.

4. *Deniability – lower bound on code parameters (Thm 4)*: Surprisingly, for any deniable code, there is also a *lower* bound on its number of codewords (as a function of the code’s structural properties). This is because of the following – suppose Alice only uses “somewhat high-weight” codewords (say all her codewords are of weight $\Omega(n^{\epsilon})$ for some $\epsilon > 0$). Then if Willie uses an analogue of minimum distance decoding he can accurately estimate Alice’s transmission status (*even* if he cannot reconstruct her message).

The first two results above are analogues (for the scenario of the BSCs considered in this work) of theorems in recent work that motivated our work (in particular, the corresponding results for Additive White Gaussian Noise (AWGN) channels proved in [1, 2]). The last two corresponding to construction of “reliable and deniable public codes”, and novel bounds on the structure of any such codes, are entirely new.

In particular, we stress again that in our model (unlike most prior work) everything that Bob knows *a priori* about Alice’s communication scheme, Willie *also* knows – there is no common randomness that is hidden from Willie that Alice and Bob can leverage. The only asymmetry between Bob’s and Willie’s estimation abilities arises from the fact that Willie’s observations of Alice’s (possible) transmissions are noisier than Bob’s. Hence the fact that we demonstrate the existence of *public* codes satisfying Result 3 above is a significant strengthening of the model in [1, 2], wherein common randomness is required, and consumed at a rate greater than the throughput of the reliable/deniable communication!

This work was partially supported by a grant from University Grants Committee of the Hong Kong Special Administrative Region, China (Project No. AoE/E-02/08), one from the CUHK MoE-Microsoft Key Laboratory for Human-centric Computing & Interface Technologies, and an SHIAE grant.

[1] If Willies channel is as good as Bobs, then no communication that is simultaneously reliable and deniable is possible, since Willie can use whatever decoding strategy Bob does.

[2] Note that this implies that Alice’s rate decays to 0 asymptotically in n. Hence in this work we usually scale Alice’s “throughput” (the number of bits in her message) with respect to $\sqrt{n}$, to obtain a quantity we call the “relative throughput”.

Also, in our model (and also in the model of [1, 2], but *not* in the vast majority of steganographic models), Alice's default transmission if she has nothing to say, is nothing. This default silence of Alice makes it challenging to hide the fact that she is *not* silent when she actually has something to say. The only reason we are able to achieve a non-zero throughput is due to the fact that Willie's observations of Alice's potential transmissions are noisy (and in particular, significantly noisier than Bob's). Hence the subtitle of this work – "hiding messages in noise". Result 4 is also novel. Similar results do not hold in the setting with common randomness between Alice and Bob – in that model (for instance that of [1, 2]), high deniability does not impose a *lower* bound on the rate of communication.

II. Related work

A. Steganography

Most steganographic models make at least one of the following assumptions (*none of which we make*):

(A1) Non-zero covertext/stegotext: In almost all works in the literature, Alice's *default* transmission (even if she has no hidden message to transmit to Bob) is usually non-zero, by this assumption. Many works characterize the capacity of various steganographic problems – see for instance [3–7]. An important exception to the non-zero covertext assumption occurs in the work of Bash, Goeckel and Towsley [1, 2] – we discuss this work in depth below.

(A2) Shared secret key/common randomness: Many steganographic protocols require a key (that is often almost as large as the message being communicated) that is shared between Alice and Bob, and is kept secret from Willie, in advance of any communication. A variety of examples of such protocols can be found in, for example [8] or [3]. However, not all works make this assumption. Some exceptions to this assumption of a shared secret include works by [4, 7].

(A3) Noiseless communication: Some works consider a model wherein the communication channel between Alice and Bob is noiseless. In some such scenarios, the optimal throughput can sometimes be boosted by a multiplicative factor of $\log n$ (for instance [8, Chapters 8 and 13]). Some model do consider noise – for instance due to an actively jamming warden (for instance [3]). In other models this may simply be random channel noise (for instance the work of [1, 2], and our work here).

B. The Square Root Law

The "Square Root Law" (often abbreviated as SRL in the literature [5, 9, 10]) can be perhaps characterized as an observation that in a variety of steganographic models, the throughput (the length of the message that Alice can communicate deniably and reliably with Bob) scales as $\mathcal{O}(\sqrt{n})$ (here n is the number of "channel uses" that Alice has access to).

We note that in our setting (and also that of [1, 2]), our throughput does indeed provably scale as the square-root of the number of channel uses. However, the critical reason underlying this scaling is that we consider the scenario wherein the covertext is all-zero – Alice must "whisper very softly", since she has no excuse if Willie hears something that cannot be explained by the noise on the channel to him.

C. The work of Bash, Goeckel and Towsley

The results and techniques closest to those in this work (and indeed the starting-point of our investigations) are those of [1, 2]. However, there are important differences in the models.

• **Public codes vs. shared secret keys:** The critical difference between our model and that of [1, 2] is that in our setting there is no shared secret key between Alice and Bob that is hidden from Willie. Hence our codes are "public". A setting wherein Alice's consumption of secret keys happens significantly faster ($\Omega(n)$) than her throughput ($\mathcal{O}(\sqrt{n})$) to Bob (as in [1, 2]) is not sustainable. Proving this required novel and powerful techniques that might be of independent interest.

• **Discrete vs. continuous channels:** In our work all channels are discrete (finite input and output alphabets) – in particular, for ease of presentation we focus here on the case wherein Alice's transmissions pass through independent BSCs. In contrast, the results of [1, 2] are for channels wherein the noise is AWGN.[3]

III. Communication Model

The transmitter Alice is connected via a binary-input binary-output broadcast medium to the receiver Bob and the warden Willie. The channels from Alice to Bob, and from Alice to Willie are respectively BSC(p_b) and BSC(p_w). By assumption, the noise on the two channels is independent, $p_b < p_w$, and Alice, Bob and Willie all know the parameters p_b and p_w.

Alice (potentially) wishes to communicate a *message* m from a set $\{1, \ldots, N\}$ to Bob – $\mathbf{M}$ denotes the random variable corresponding to m. (If Alice is not transmitting, her message is 0.) If Alice *does* wish to communicate with Bob, then a certain (arbitrary) binary variable $\mathbf{T} \triangleq 1$, else $\mathbf{T} \triangleq 0$. Only Alice knows the value of $\mathbf{T}$ *a priori*.

Alice encodes each message m into a length-n binary *codeword* $\vec{\mathbf{x}}_m$ using an *encoder* $Enc(\cdot) : \{0\} \cup \{1, \ldots, N\} \rightarrow \{0,1\}^n$. The encoder always maps the 0 message to the zero-vector $\vec{\mathbf{0}}$. The set $\{\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_N\}$ of possible non-zero codewords is denoted by a specific *codebook* $\mathcal{C}_0$. The *throughput* τ of Alice's codebook is defined as $\log N$, and the *relative throughput* r of Alice's codebook is defined as $\log N/\sqrt{n}$.[4]

Bob receives the length-n binary vector $\vec{\mathbf{Y}}_b = \vec{\mathbf{X}} \oplus \vec{\mathbf{Z}}_b$, where $\vec{\mathbf{Z}}_b$ denotes BSC(p_b) channel noise. Bob uses his *decoder* $Dec(\cdot) : \{0,1\}^n \rightarrow \{0\} \cup \{1, \ldots, N\}$ to generate his *estimate of Alice's message* $\hat{\mathbf{M}} = Dec(\vec{\mathbf{Y}}_b)$. Bob's *error probability* is defined as $\Pr_{\mathbf{M}, \vec{\mathbf{Z}}_b}(\hat{\mathbf{M}} \neq \mathbf{M})$. Alice's codebook $\mathcal{C}_0$ is said to be $(1-\epsilon)$-*reliable* if Bob's error probability is less than ϵ.

[3] It is conceivable that our proof techniques also carries over to the AWGN model of [1, 2], but significant extensions would be required to translate our techniques from the discrete world over to the continuous version.

[4] In this work, the throughput of the codes we consider typically scales with the block-length n as $\mathcal{O}(\sqrt{n})$. This corresponds to a "rate" going to 0 as n goes to ∞, rather than converging to a constant as is common in many other communication settings. Thus we consider the relative throughput instead of the rate of our codes.

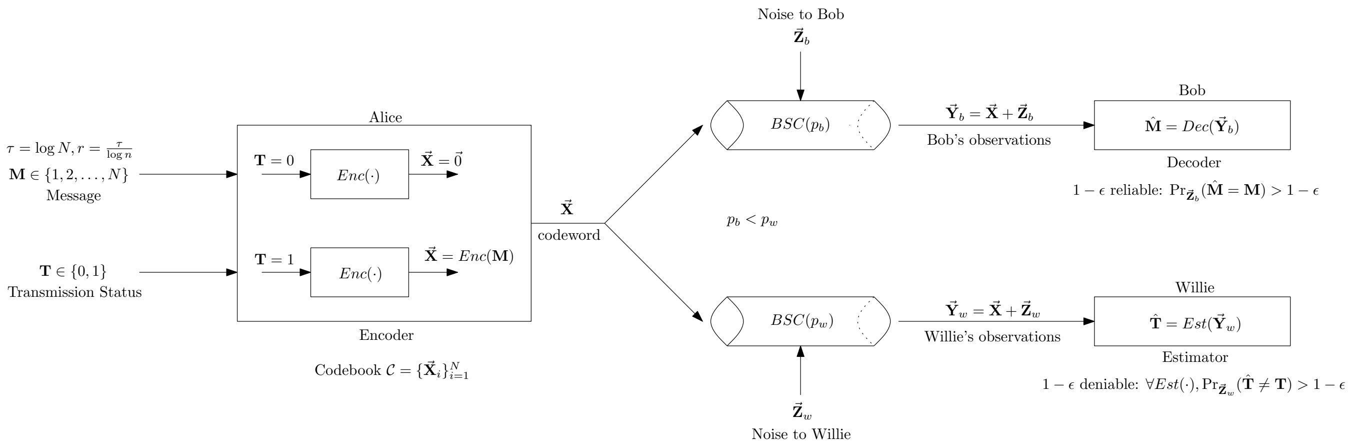
Noise to Bob
$\vec{\mathbf{Z}}_b$
Alice
$\tau = \log N, r = \frac{\tau}{\log n}$
$\mathbf{M} \in \{1, 2, \ldots, N\}$
Message
$\mathbf{T} = 0$
$Enc(\cdot)$
$\vec{\mathbf{X}} = \vec{0}$
$\mathbf{T} \in \{0, 1\}$
Transmission Status
$\mathbf{T} = 1$
$Enc(\cdot)$
$\vec{\mathbf{X}} = Enc(\mathbf{M})$
$\vec{\mathbf{X}}$
codeword
Encoder
Codebook $\mathcal{C} = \{\vec{\mathbf{X}}_i\}_{i=1}^N$
$BSC(p_b)$
$\vec{\mathbf{Y}}_b = \vec{\mathbf{X}} + \vec{\mathbf{Z}}_b$
Bob's observations
Bob
$\hat{\mathbf{M}} = Dec(\vec{\mathbf{Y}}_b)$
Decoder
$1 - \epsilon$ reliable: $\Pr_{\vec{\mathbf{Z}}_b}(\hat{\mathbf{M}} = \mathbf{M}) > 1 - \epsilon$
$p_b < p_w$
$BSC(p_w)$
$\vec{\mathbf{Y}}_w = \vec{\mathbf{X}} + \vec{\mathbf{Z}}_w$
Willie's observations
Willie
$\hat{\mathbf{T}} = Est(\vec{\mathbf{Y}}_w)$
Estimator
$1 - \epsilon$ deniable: $\forall Est(\cdot), \Pr_{\vec{\mathbf{Z}}_w}(\hat{\mathbf{T}} \neq \mathbf{T}) > 1 - \epsilon$
$\vec{\mathbf{Z}}_w$
Noise to Willie

Fig. 1. Depending on her transmission status $\mathbf{T}$, Alice either broadcasts the all-zero dector $\vec{\mathbf{0}}$, or encodes her messages $\mathbf{M}$ into codewords $\vec{\mathbf{X}} = Enc(\mathbf{M})$ from her codebook $\mathcal{C}$. The codebook has $2^\tau = 2^{r\sqrt{n}}$ codewords, where the throughput τ scales as $\mathcal{O}(\sqrt{n})$. Bob receives $\vec{\mathbf{Y}}_b$ over a BSC(p_b), and Willie receives a noisier version $\vec{\mathbf{Y}}_w$ over a BSC(p_w), where $p_b < p_w$. It is desired that the codebook $\mathcal{C}$ be both reliable (*i.e.* Bob is able to decode $\mathbf{M}$ as $\hat{\mathbf{M}}$ with a "small" probability error) and deniable (*i.e.* Willie's observations give him essentially no information about Alice's transmission status.)

Willie knows *a priori* both $Enc(\cdot)$ and $Dec(\cdot)$. Willie receives the length-n binary vector $\vec{\mathbf{Y}}_w = \vec{\mathbf{X}} \oplus \vec{\mathbf{Z}}_w$, where $\vec{\mathbf{Z}}_w$ denotes the BSC(p_w) channel noise. Willie uses his *estimator* $Est_{\mathcal{C}_0}(\cdot) : \{0,1\}^n \rightarrow \{0,1\}$ to generate his *estimate of Alice's transmission status* as $\hat{\mathbf{T}} = Est_{\mathcal{C}_0}(\vec{\mathbf{Y}}_w)$.

We use a hypothesis-testing metric to quantify the *deniability of Alice's code*. Let *the probability of false alarm* $\Pr_{\vec{\mathbf{Z}}_w}(\hat{\mathbf{T}} = 1 | \mathbf{T} = 0)$ be denoted by $\alpha(Est_{\mathcal{C}_0}(\cdot))$. Analogously, let *the probability of missed detection* $\Pr_{\mathbf{M},\vec{\mathbf{Z}}_w}(\hat{\mathbf{T}} = 0 | \mathbf{T} = 1)$ be denoted by $\beta(Est_{\mathcal{C}_0}(\cdot))$ These quantities denote respectively the probabilities that Willie guesses Alice is transmitting even if she is not, and that Willie guesses Alice is not transmitting even though she is. We say Alice's codebook $\mathcal{C}_0$ is $(1-\epsilon)$*-deniable* if there is no estimator $Est_{\mathcal{C}_0}(\cdot)$ such that $\alpha(Est_{\mathcal{C}_0}(\cdot)) + \beta(Est_{\mathcal{C}_0}(\cdot)) < 1 - \epsilon$.We henceforth denote $\alpha(Est_{\mathcal{C}_0}(\cdot))$ and $\beta(Est_{\mathcal{C}_0}(\cdot))$ simply by α and β.

For any block-length n, we say a corresponding codebook $\mathcal{C}_0$ is *simultaneously* $(1-\epsilon)$*-reliable and* $(1-\epsilon)$*-deniable* if it simultaneously ensures that Bob's probability of decoding error is at most ϵ, and has deniability $1 - \epsilon$.

IV. Main results/high-level intuition

We present our main results and the underlying intuition. **Due to lack of space we do not present the proof details, and instead refer the interested reader to [11].**

Theorem 1 below proves an outer bound on the fraction of "high Hamming weight" codewords of any highly deniably $\mathcal{C}$. The intuition is that if many codewords have "high" Hamming weight, with significant probability the Hamming weight of Willie's observed vector $\vec{\mathbf{y}}_w$ is above a certain threshold.

Theorem 1. *For any $p_w < 1/2$, if more than a γ fraction of the codewords in $\mathcal{C}$ are of weight greater than $c_1\sqrt{n}$, then $\mathcal{C}$ is less than $\left(1 - \gamma + \frac{8\gamma p_w(1-p_w)}{c_1^2(1-2p_w)^2}\right)$-deniable.*

Theorem 2 below uses information-theoretic converse techniques, along with Theorem 1), to prove an upper bound on the throughput τ of any code that simultaneously has high reliability and deniability.

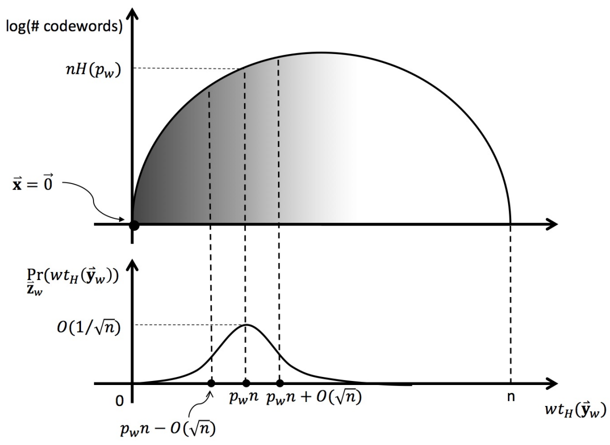

Fig. 2. **Willie's observation if Alice does not transmit:** The upper curve represents the set of all possible $\vec{\mathbf{y}}_w$ that Willie may observe if Alice transmits nothing. The $\vec{\mathbf{y}}_w$ are arranged so that vectors with lower Hamming weight are to the left of vectors with higher Hamming weight, and the height of the enclosing curve (the binary entropy function) denotes the (logarithm of the) number of binary vectors of a particular Hamming weight. Hence the shaded region denotes the set of "likely" $\vec{\mathbf{y}}_w$ that Willie observes, with the density denoting probability of observing corresponding $\vec{\mathbf{y}}_w$s. The curve at the bottom plots the probability distribution of observing $\vec{\mathbf{y}}_w$ of a particular Hamming weight. Since Alice's transmitted codeword is $\vec{\mathbf{0}}$, the "typical" $\vec{\mathbf{y}}_w$ that Willie observes are of weight approximately $p_w n$ (with a variation of $\mathcal{O}(\sqrt{n})$). This curve is "smooth" and follows a binomial distribution.

Theorem 2. *For any sufficiently small ϵ, if a codebook $\mathcal{C}$ is simultaneously $(1-\epsilon)$-deniable and $(1-\epsilon)$-reliable, $p_w > p_b$, and $p_b \in (0,1)$ the relative throughput r is at most*

$$\sqrt{p_w(1-p_w)}\left(\frac{1-2p_b}{1-2p_w}\right)(1-2\epsilon)^{-3/2}\log\left(\frac{1-p_b}{p_b}\right).$$

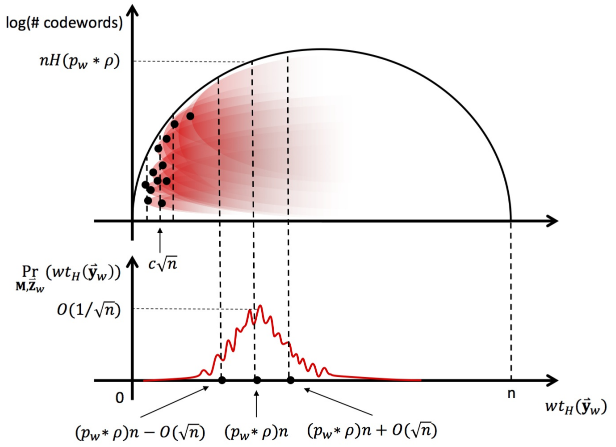
log(# codewords)
$nH(p_w * \rho)$
$c\sqrt{n}$
$\Pr_{\mathbf{M},\vec{\mathbf{Z}}_w}(wt_H(\vec{\mathbf{y}}_w))$
$O(1/\sqrt{n})$
0
n
$wt_H(\vec{\mathbf{y}}_w)$
$(p_w * \rho)n - O(\sqrt{n})$
$(p_w * \rho)n$
$(p_w * \rho)n + O(\sqrt{n})$

Fig. 3. **Willie's observations if Alice transmits:** The red region denotes the set of $\vec{\mathbf{y}}_w$ that Willie may observe if Alice transmits a codeword. The black dots on the left denote codewords of $\mathcal{C}$. If Alice transmits a particular $\vec{\mathbf{x}}$, the set of $\vec{\mathbf{y}}_w$ that Willie is likely to observe is shown by the red paraboloid region extending rightwards from that $\vec{\mathbf{x}}$. The overall probability distribution over Willie's observed $\vec{\mathbf{y}}_w$ is hence the "average" of the paraboloid regions. In this case the probability distribution on $\vec{\mathbf{y}}_w$ is somewhat "lumpy", since the probability that Willie observes a particular $\vec{\mathbf{y}}_w$ depends on the distribution of the Hamming distance between that particular $\vec{\mathbf{y}}_w$ and the set of codewords $\vec{\mathbf{x}} \in \mathcal{C}$. So the weight distribution of $\vec{\mathbf{y}}_w$ is a weighted sum of binomial distributions.

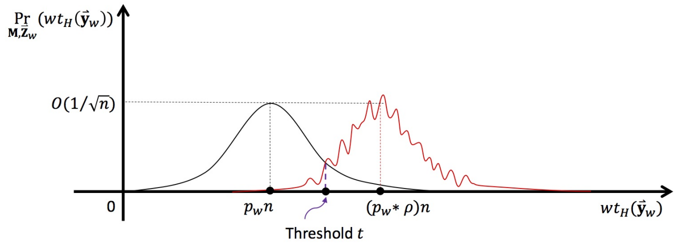

Fig. 4. **A threshold estimator for Willie:**If "too many" codewords in the codebook $\mathcal{C}$ have "large" Hamming weight, the two probability distributions on $\vec{\mathbf{y}}_w$ corresponding to $\mathbf{T} = 0$ and $\mathbf{T} = 1$, respectively $\Pr_{\vec{\mathbf{Z}}_w}(\vec{\mathbf{y}}_w|\mathbf{T} = 0)$ and $\Pr_{\mathbf{M},\vec{\mathbf{Z}}_w}(\vec{\mathbf{y}}_w|\mathbf{T} = 1)$, are "very different". In this case, Willie can detect Alice's transmission status fairly accurately by using the following "thereshold" estimator. In particular, Willie's estimator outputs $\hat{\mathbf{T}} = 1$ if the weight of the observed codeword $\vec{\mathbf{y}}_w$ is above a carefully chosen threshold, and $\hat{\mathbf{T}} = 0$ otherwise. This figure visually depicts Theorem 1.

Next we state, and sketch the proof of, one of the main results of this work – namely, that randomly chosen codes (chosen from a suitable ensemble) are with high probability simultaneously highly reliable and highly deniable.

The code constructions in [1, 2] use this common randomness as follows. First, the common randomness is used to coordinate which of an ensemble of possible codebooks Alice actually uses to communicate with Bob. Since the common randomness is kept secret from Willie, if Alice transmits a non-zero codeword, the probability distribution on his received $\vec{\mathbf{y}}_w$ is that of a *code ensemble average* $\Pr_{\mathcal{C},\mathbf{M},\vec{\mathbf{Z}}_w}(\vec{\mathbf{y}}_w|\mathbf{T} = 1)$. Using some elegant statistical properties of this ensemble average distribution, and the distribution $\Pr_{\vec{\mathbf{Z}}_w}(\vec{\mathbf{y}}_w|\mathbf{T} = 0)$ (corresponding to Willie's observations if Alice does not transmit anything), [1, 2] demonstrate that Willie is essentially unable to learn anything about the binary random variable $\mathbf{T}$, since the two distributions "look very similar" from Willie's observations of $\vec{\mathbf{y}}_w$. Their proof can be essentially summarized in the statement that "the ensemble average codebook is highly deniable". The challenge in extending their proof technique to a public codebook is that this proof says nothing about the existence of a single, public, highly deniable codebook.

Our key idea is to extend the analysis by proving that the *actual* distribution $\Pr_{\mathbf{M},\vec{\mathbf{Z}}_w}(\vec{\mathbf{y}}_w|\mathbf{T} = 1)$ of $\vec{\mathbf{y}}_w$ if Alice transmits a non-zero codeword is tightly concentrated about its ensemble average. However, our first "naïve" attempts in using standard concentration inequalities were unsuccessful, since for any particular $\vec{\mathbf{y}}_w$ the probability (averaged over

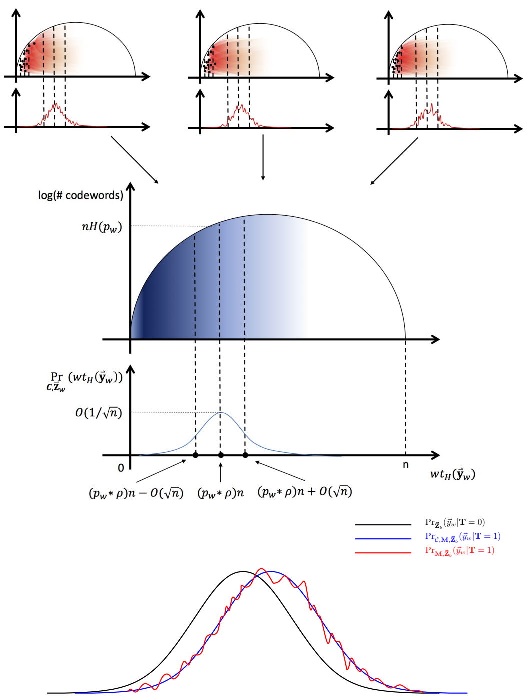
log(# codewords)
$nH(p_w)$
$\Pr_{\mathcal{C},\vec{\mathbf{Z}}_w}(wt_H(\vec{\mathbf{y}}_w))$
$O(1/\sqrt{n})$
0
n
$wt_H(\vec{\mathbf{y}}_w)$
$(p_w * \rho)n - O(\sqrt{n})$
$(p_w * \rho)n$
$(p_w * \rho)n + O(\sqrt{n})$
$\Pr_{\vec{\mathbf{Z}}_b}(\vec{y}_w|\mathbf{T} = 0)$
$\Pr_{\mathcal{C},\mathbf{M},\vec{\mathbf{Z}}_b}(\vec{y}_w|\mathbf{T} = 1)$
$\Pr_{\mathbf{M},\vec{\mathbf{Z}}_b}(\vec{y}_w|\mathbf{T} = 1)$

Fig. 5. **Deniability from Willie:** Our proof that a random codebook $\mathcal{C}$ chosen with the "right" parameters (number of codewords, expected weight of codewords) proceeds as follows. We need to demonstrate that the probability distributions $\Pr_{\vec{\mathbf{Z}}_w}(\vec{\mathbf{y}}_w|\mathbf{T} = 0)$ and $\Pr_{\mathbf{M},\vec{\mathbf{Z}}_w}(\vec{\mathbf{y}}_w|\mathbf{T} = 1)$ are "close" (in variational distance). However, since the latter distribution is complex (due to its dependence on the specific codebook $\mathcal{C}$), we do this comparison in two stages. We first compute the *ensemble distribution* of $\vec{\mathbf{y}}_w$, *i.e.*, the "smooth blue" region/curve denoting the "ensemble average" (over all suitably chosen random codebooks) of the probability distribution on $\vec{\mathbf{y}}_w$. We then demonstrate that the probability distribution $\Pr_{\vec{\mathbf{Z}}_w}(\vec{\mathbf{y}}_w|\mathbf{T} = 0)$ and the ensemble distribution $\Pr_{\mathcal{C},\mathbf{M},\vec{\mathbf{Z}}_w}(\vec{\mathbf{y}}_w|\mathbf{T} = 1)$ (*i.e.* the weighted average over all possible codebooks $\mathcal{C}$ of the latter distribution) are "close". Finally, we prove that with high probability over the choice of codebooks $\mathcal{C}$, the distribution of $\Pr_{\mathbf{M},\vec{\mathbf{Z}}_w}(\vec{\mathbf{y}}_w|\mathbf{T} = 1)$ is tightly concentrated around its expectation $\Pr_{\mathcal{C},\mathbf{M},\vec{\mathbf{Z}}_w}(\vec{\mathbf{y}}_w|\mathbf{T} = 1)$. This figure visually depicts deniability in Theorem 3.

Alice's choice of message, channel noise, and over all codebooks) that Willie actually observes $\vec{\mathbf{y}}_w$ is exceedingly small (decaying at least exponentially in n). This means that standard concentration inequalities such as the Chernoff bound fail to give the required probability of concentration. This is because in these bounds the probability of concentration depends on the expected values over $\mathbf{M}$, $\vec{\mathbf{Z}}_w$ and $\mathcal{C}$ of observing a particular $\vec{\mathbf{y}}_w$ – these expected values are "too small". Hence we proceed indirectly. We first note that it suffices to prove that $\Pr_{\mathbf{M},\vec{\mathbf{Z}}_w}(\vec{\mathbf{y}}_w|\mathbf{T}=1)$ converges point-wise to its ensemble average for "typical" $\vec{\mathbf{y}}_w$ (since the bulk of the probability mass of the ensemble average distribution falls in a certain range). For any $\vec{\mathbf{y}}_w$ in this range, we prove that expected number of codewords at a certain distance range (corresponding to the "typical" noise patterns $\vec{\mathbf{Z}}_w$) of each $\vec{\mathbf{y}}_w$ is super-polynomial. For random variables with such "large" (super-polynomial) expectations, standard arguments suffice to prove concentration with probability that is super-exponentially small in n. This allows us to show that with high probability over the ensemble average, a randomly chosen codebook satisfies the property that the number of codewords in "typical" Hamming shells around most "typical" $\vec{\mathbf{y}}_w$ are tightly concentrated around their expectations. Book-keeping calculations then enable us to show that this concentration in the distance-distribution of codewords translate to a point-wise concentration (with super-exponential probability) of $\Pr_{\mathbf{M},\vec{\mathbf{Z}}_w}(\vec{\mathbf{y}}_w|\mathbf{T}=1)$ to its ensemble average. This technique allows us to bypass the problem of the small expected values [5] of the random variables of primary interest (the probability of observing specific channel outputs if a specific codebook is used), by focusing instead on random variables with "large expected values" [6] (numbers of codewords of certain "types") that then enable us to recover the random variables of primary interest. One calculation that requires some care is that due to the low throughput of our codes (scaling as $\mathcal{O}(\sqrt{n})$) we need to define our typical sets carefully, to simultaneously ensure that they are high probability sets, but are also not "too large".

To complete the proof, we need to demonstrate that a randomly chosen code is also highly reliable with sufficiently high probability, and hence a randomly chosen code is, w.h.p., simultaneously high deniable and highly reliable. This follows from somewhat standard random coding arguments, if Bob decodes to the nearest codeword. One calculation that requires some care is due to the low expected weight of codewords (by construction chosen to be about $\mathcal{O}(\sqrt{n})$), and hence the notion of "typicality decoding" has to be carefully defined.

We generate the ensemble of all codebooks $\mathcal{C}$ containing $2^{r\sqrt{n}}$ codewords of block-length n, with each codeword generated by choosing each bit to be 1 with probability Bernoulli(ρ), and sample codebooks from this distribution (we call such codebooks *random public codebooks*). In what follows, we set ρ to equal $c_2/\sqrt{n}$, where c_2 is a code design parameter.

Let $c_8 = 2\sqrt{\frac{\frac{3}{\sqrt{2}}\sqrt{\ln\frac{2}{\epsilon_2}}+\epsilon_1\sqrt{\ln 2}}{\epsilon_1\sqrt{\ln 2}(1-\delta_0)}}$, $c_9 = 4(1-2p_w)^2\left(\sqrt{\frac{1}{p_w}+\frac{1}{1-p_w}}\sqrt{\ln\frac{8}{\epsilon}}+\frac{\epsilon}{4}\sqrt{\ln 2}\right)$, and $c_{10} = \frac{\epsilon}{4}\sqrt{\ln 2}(1-2p_b)^2(1-\delta_0)$. where δ_0 is the positive root of the quadratic equation $\frac{1}{2}\delta^2 - 2(\frac{1}{2}-p_b)^2(1-\delta) = 0$.

Theorem 3. *For any $p_w > \max\{\frac{1}{3}, \frac{1}{2} - \frac{1-2p_b}{2c_8}\}$, and any relative throughput $r \in (c_9, c_{10})$ with probability greater than $1-2^{-\theta(\sqrt{n})}$, a random public codebook $\mathcal{C}_0$ is simultaneously $(1-\epsilon)$-reliable and $(1-\epsilon)$-deniable.*

Finally, we are able to use our proof techniques to prove a novel lower bound on the throughput of any highly deniable code. Intuitively, our argument follows from noting that the set of the $\vec{\mathbf{y}}_w$ that are "noise-typical" with respect to the all-zero codeword (corresponding to Alice not transmitting anything) comprises of $\vec{\mathbf{y}}_w$ sequences of weight approximately $p_w n$. But any transmitted non-zero codeword $\vec{\mathbf{x}}$ can only be "noise-typical" with respect to a subset of these $\vec{\mathbf{y}}_w$ sequences of weight approximately $p_w n$. But for $\Pr_{\vec{\mathbf{Z}}_w}(\vec{\mathbf{y}}_w|\mathbf{T}=0)$ and $\Pr_{\mathbf{M},\vec{\mathbf{Z}}_w}(\vec{\mathbf{y}}_w|\mathbf{T}=1)$ to be "close" (so that the code is highly deniable), the set of $\vec{\mathbf{y}}_w$ that are "noise-typical" with respect to some transmitted codeword should "cover" most of the sequences of weight approximately $p_w n$. We then use a counting argument to bound from *below* a function of the weight distribution of any highly-deniably codebook.

For each $i \in \{1,\ldots,n\}$, let $N_i(\mathcal{C})$ denote the number codewords of Hamming weight i in Alice's codebook $\mathcal{C}$.

Theorem 4. *For any code $\mathcal{C}$ that is $(1-\epsilon)$-deniable,*

$$1-\epsilon \le \frac{1}{\sum_{i=1}^{n} N_i(\mathcal{C})} \sum_{i=0}^{n} \left[(N_i(\mathcal{C})+1)e^{-2(\frac{1}{2}-p_w)^2 i}\right] N_i(\mathcal{C}).$$

This shows, for instance, that if one chooses wishes to choose a highly deniable codebook such that *all* codewords are of weight about $\sqrt{n}$, then, in fact, there is a *lower* bound on the number of codewords that scales as $2^{\Omega(\sqrt{n})}$.

References

[1] B. Bash, D. Goeckel, and D. Towsley, "Square root law for communication with low probability of detection on AWGN channels," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 2012, pp. 448–452.

[2] ——, "Limits of reliable communication with low probability of detection on AWGN channels," *arXiv preprint arXiv:1202.6423*, 2012.

[3] Y. Wang and P. Moulin, "Perfectly secure steganography: Capacity, error exponents, and code constructions," *IEEE Transactions on Information Theory*, vol. 54, no. 6, pp. 2706–2722, 2008.

[4] B. Ryabko and D. Ryabko, "Asymptotically optimal perfect steganographic systems," *Problems of Information Transmission*, vol. 45, no. 2, pp. 184–190, 2009.

[5] A. D. Ker, "The square root law in stegosystems with imperfect information," in *Proceedings of the 12th Information Hiding Workshop*, 2010.

[6] ——, "The square root law requires a linear key," in *in Proc. 11th ACM Workshop on Multimedia and Security*, 2009, pp. 85–92.

[7] A. Ker, "The square root law does not require a linear key," in *Proceedings of the 12th ACM workshop on Multimedia and security*, 2010, pp. 213–224.

[8] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge University Press, 2009.

[9] A. Ker, T. Pevný, J. Kodovský, and J. Fridrich, "The square root law of steganographic capacity," in *Proceedings of the 10th ACM workshop on Multimedia and security*. ACM, 2008, pp. 107–116.

[10] A. Ker, "A capacity result for batch steganography," *Signal Processing Letters, IEEE*, vol. 14, no. 8, pp. 525–528, 2007.

[11] P. H. Che, M. Bakshi and S. Jaggi, "Reliable deniable communication: Hiding messages in noise", Project website: http://personal.ie.cuhk.edu.hk/~cph010/project-steganography.html

[5] Or, as George Walker Bush put it, "the soft bigotry of low expectations".

[6] Or, as Philip Pirrip might put it, we have "Great Expectations".