RUHR-UNIVERSITÄT BOCHUM
Horst Görtz Institute for IT Security

**Technical Report TR-HGI-2016-001**

# SkypeLine
# Robust Hidden Data Transmission for VoIP

*Katharina Kohls, Thorsten Holz, Dorothea Kolossa, Christina Pöpper*

Group for Information Security

Ruhr-Universität Bochum
Horst Görtz Institute for IT Security
D-44780 Bochum, Germany

TR-HGI-2016-001
February 25, 2016

RUHR UNIVERSITÄT BOCHUM RUB

hgi Horst Görtz Institut für IT-Sicherheit

# SkypeLine
# Robust Hidden Data Transmission for VoIP

Katharina Kohls, Thorsten Holz, Dorothea Kolossa, Christina Pöpper

**Abstract**

Internet censorship is used in many parts of the world to prohibit free access to online information. Different techniques such as IP address or URL blocking, DNS hijacking, or deep packet inspection are used to block access to specific content on the Internet. In response, several censorship circumvention systems were proposed that attempt to bypass existing filters. Especially systems that hide the communication in different types of cover protocols attracted a lot of attention. However, recent research results suggest that this kind of covert traffic can be easily detected by censors.

In this paper, we present *SkypeLine*, a censorship circumvention system that leverages Direct-Sequence Spread Spectrum (DSSS) based steganography to hide information in Voice-over-IP (VoIP) communication. SkypeLine introduces two novel modulation techniques that hide data by modulating information bits on the voice carrier signal using pseudo-random, orthogonal noise sequences and repeating the spreading operation several times. Our design goals focus on undetectability in presence of a strong adversary and improved data rates. As a result, the hiding is inconspicuous, does not alter the statistical characteristics of the carrier signal, and is robust against alterations of the transmitted packets. We demonstrate the performance of SkypeLine based on two simulation studies that cover the theoretical performance and robustness. Our measurements demonstrate that the data rates achieved with our techniques substantially exceed existing DSSS approaches. Furthermore, we prove the real-world applicability of the presented system with an exemplary prototype for Skype.

## 1 Introduction

In different parts of the world, Internet Service Providers (ISPs) perform censorship to filter certain content and prohibit users from accessing specific information. Such censorship is often performed in repressive regimes, where the government wants to prevent citizens from freely accessing the Internet. From a technical point of view, these censorship systems are based on methods such as IP address or URL blocking, DNS hijacking, or deep packet inspection to block specific content. We can expect that further censorship systems will be developed in the future that leverage increasingly sophisticated methods for blocking content.

In response, different kinds of censorship circumvention systems were developed in the last years that enable unrestricted access to the Internet [11, 16, 18, 25]. For example, anonymous communication systems like Tor [11] utilize encrypted tunnels between different proxies to provide sender anonymity. Unfortunately, such systems are vulnerable to attacks where the censors enumerate all proxies and block access to them. In response to such attacks, so-called private bridges [11] were introduced, but the actual communication messages can be detected on the network layer and empirical data suggest that repressive regimes are still capable of blocking Tor [23]. The root cause lies in the fact that anonymous communication systems leverage protocol messages that can be fingerprinted on the network layer and subsequently are blocked.

To overcome such problems, several steganographic systems based on the idea of hiding the communication in different kinds of cover protocols were proposed [17, 26, 28, 29, 39, 40]. Those systems have in common that they embed the hidden messages in cover traffic that censors cannot block in practice since such blocking would have significant side-effects (so-called *unblockability* [25]). For example, SkypeMorph [29] mimics Skype video call messages by constructing

messages in such a way that the packet sizes and sending times are similar to benign video calls. FreeWave [17] takes this idea one step further and proposes to send IP traffic over Skype voice calls using a virtual modem ("IP over VoIP"). In both cases, the resulting communication messages mimic legitimate Skype messages.

However, the underlying idea of providing *unobservability* [25] (i.e., a censor cannot decide whether or not a given client uses a censorship circumvention system) is not achieved by such approaches. Houmansadr et al. [15] demonstrate that accurately mimicking a cover protocol is very challenging in practice and they introduce different attacks to distinguish covert traffic from benign traffic. Furthermore, Geddes et al. [13] showed that censors can easily detect and disrupt the covert communication channels provided by systems such as SkypeMorph and FreeWave by actively interfering with the protocols. A major problem is that approaches like SkypeMorph and FreeWave are vulnerable to eavesdropping attacks where a censor listens to the communication: since no actual (human) voice messages are exchanged, the censor only hears noise and can thus easily spot the covert channel. We expect that such attacks on the endpoint are actually viable in practice: TOM-Skype, a modified version of Skype in China, already censors and surveils text chat [22] and this can be easily extended to also cover voice calls, thus enabling a censor to fully eavesdrop calls.

In this paper, we present SkypeLine, a novel steganographic censorship circumvention system based on the idea of hiding secret data by embedding it in a VoIP system using Direct-Sequence Spread Spectrum (DSSS) modulation: our approach spreads information bits by DSSS and adds them to the audio signal before the VoIP encoding and packaging. This happens in such a way that only parties in possession of a pre-shared secret can detect the hidden communication and decode the information bits. By hiding the covert messages directly within the audio signal, it becomes a challenging problem to actually spot covert messages: we avoid architectural, channel, and content mismatches, as described in [13], which are the building blocks for the security of steganographic systems.

In the design of our solution, we need to address the following technical challenges: First, VoIP tools typically utilize connectionless transmission protocols for low latencies at the expense of packet loss. This is acceptable for multimedia data, but can interfere with loss-sensitive secret data. Second, standard DSSS-based approaches often achieve only very small data rates and it is desirable to increase their throughput. Third, audio codecs of VoIP systems are trained to optimize and compress the audio input and contained noise when creating the digital data stream. This can disrupt the successful recovery of hidden information.

We analyze the performance of SkypeLine with respect to our two main design goals, namely security in presence of a strong eavesdropping adversary and improved, robust throughput rates. To this end, we conducted theoretical and practical measurements and demonstrate that, compared to previous DSSS-based work, our approach significantly increases the data rate. At the same time, it provides the potential for statistically and acoustically unobservable covert communication. We can provide security in a most restrictive censorship scenario while reliable data transmission is assured. Within this highly demanding deployment scenario more performant hiding schemes cannot be applied.

In summary, we make the following three contributions:

- We propose *SkypeLine*, a novel steganography system that provides the transmission of secret data through a VoIP carrier system based on DSSS. It can be deployed with arbitrary VoIP clients and introduces two novel modulation techniques, namely one *parallel binary* and one *m-ary* approach that conceptually differ from existing techniques. These hiding techniques can achieve significantly increased data rates ranging from approx. 224 bps (max. binary) to 2400 bps (max. m-ary).
- We show that SkypeLine hides inconspicuous (subjectively based on hearing tests, mathematically based on PESQ[1]), does not alter the statistical characteristics of the carrier signal to a level that allows detection, and is robust against alterations of the transmitted packets. The latter feature is achieved by an optional acknowledgment mechanism that identifies and

[1] `http://www.pesq.org/`

retransmits missing or duplicate data.

- For a demonstration of the proposed scheme's capabilities, we conduct two simulation studies and real-world measurements: we i) assess and optimize the performance of the binary and m-ary modulation schemes in a MATLAB-based simulation study that reveals the system's general performance. Exemplary for ii) Skype we perform prototypical real-world measurements that demonstrate the practical feasibility of the approach. iii), in an OMNeT++ network simulation study we show how an acknowledgment scheme can overcome loss and alterations at the transmission channel.

# 2 Technical Background

In this section, we summarize the technical background of the proposed system. This includes the VoIP carrier system as well as the modulation and encoding schemes.

## 2.1 VoIP Carrier System

VoIP allows for exchanging audio and video data between communication parties over the Internet. The service is available through various clients such as [1, 2, 21] that manage the set-up and teardown of conversations, transmission of information, and optimization of multimedia data. Through its general acceptance it is a legitimate carrier for covert communication, while no additional client software or infrastructure is required.

In general the transmission of multimedia information is performed by the connectionless UDP protocol and provides real-time communication at the expense of transmission reliability. That is, lost packets will not be recovered and result in a degradation of media quality at the receiver. As the secret information is directly attached to the multimedia data, different load situations of the network can affect the reliability of a circumvention system that communicates sensitive information.

In preparation for a packet-wise transmission the digital audio input gets optimized and compressed by the VoIP client and its audio codec. Recent codecs such as Opus [3] employ vector quantization whereby an audio signal is fragmented into frames of dynamic length. For compression, each frame gets represented through a minimum distance, that is, most similar, entry in the codec's code book. In context of modulation-based circumvention schemes the activity of a codec can interfere with the successful recovery of hidden information, as it applies slight changes to the original audio signal.

## 2.2 Modulation

As we make use of DSSS for information hiding, we briefly introduce it in the following.

### 2.2.1 Direct-Sequence Spread Spectrum

When the DSSS modulation technique is applied on a carrier signal, it is directly adapted by a high-frequency spreading sequence. Originally proposed in the context of wireless signal transmissions, DSSS allows for hidden communication since the resulting wideband signal appears as a noise signal; this provides a certain level of resistance to intentional and unintentional interference with the signal transmission. Based on these characteristics, the transmission of information is generally hard to detect for an attacker and—given the spreading sequence that is specific for each communication session—also allows for sharing the transmission medium with multiple users. We profit from these characteristics for a hiding process that resists speech optimization and is hard to detect.

As shown in Figure 1, we apply a similar technique to audio signals in our approach. On the sender side, one bit of information $B \in \{0, 1\}$ is mapped to an intermediary bit $s = 2B-1 \in \{-1, 1\}$

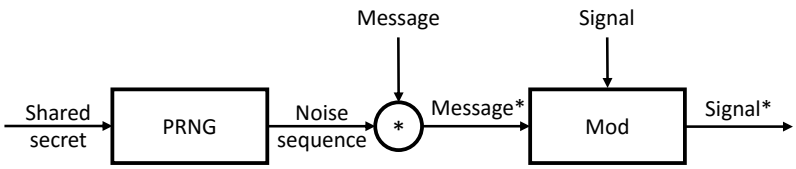


Figure 1: General hiding process with PRNG (shift register) and modulation module.

and then spread over a pseudo-noise sequence $\vec{n} = (n_1, n_2, ...n_\vartheta)^T \in \{-1, 1\}$ of length $\vartheta$:

$$\vec{n}' = \vec{n} \cdot s, \tag{1}$$

where $\vec{n}'$ is the modulated noise sequence and $\cdot$ is a scalar multiplication. After spreading the information bit, $\vec{n}'$ is added to portions of the input signal $\vec{a}$ of the same length $\vartheta$. This results in the information-carrying output $\vec{a}'$:

$$\vec{a}' = \vec{a} + \vec{n}'. \tag{2}$$

The modulation and addition is repeated for each bit of information. Therefore, the final audio signal carries multiple consecutive noise sequences that each include one individual bit of the secret message. The hidden information can be reconstructed with knowledge of the noise sequence $\vec{n}$ (see Section 4 for a detailed description).

#### 2.2.2 Pseudo Noise Generation

In our scheme, modulated pseudo-noise (PN) sequences $\vec{n}'$ are added to the audio input for attaching the secret information. The sequences should provide suitable properties for being unobtrusive to an observing attacker while being easy to modulate and demodulate. This concerns acoustic unobtrusiveness, where the attacker is unable to distinguish modulated and unmodulated signals by hearing, as well as a statistical indistinguishablility of signals.

Binary PN sequences can be generated by shift registers that output periodic sequences with specific correlation properties. The initialization of a shift register is seeded with a *shared secret* between sender and receiver that defines the output of the shift register and meta information, e.g., the length of the noise sequence. In case of White Gaussian Noise (WGN) the full noise sequence must be part of the shared secret, as it cannot be generated by a seeded shift register.

The auto-correlation (cross-correlation of the signal with itself with different offsets) of the PN sequences should be close to 0 for all offsets $\neq 0$ to enable a receiver to synchronize the demodulation operation with the sender's signal. Furthermore, good cross-correlation characteristics allow for distinguishing multiple noise candidates, which is required for the presented parallel binary and m-ary modulation techniques. Suitable correlation characteristics are provided by Gold codes and WGN sequences. We analyze the performance and robustness of the modulation scheme for different PN types, which is described in Section 4.2.

# 3 Assumptions

In general, we assume that the sender and the intended receiver(s) have established a shared secret by out-of-band channels before the start of their communication. This secret is either a seed to a shift register or a set of WGN sequences. We next present our threat model and design goals towards which the implementation has been optimized.

## 3.1 Threat Model

We assume that the communication is established in a censored area such as under the Golden Shield Project [42], where a censoring ISP regulates the communication and the extent of available

content. VoIP services can be classified as key services and they are typically available within censored areas since they can also be used in favor of the censor.

Censors are able to access all information exchanged between two communicating parties. This includes instant messages and eavesdropping of VoIP calls, which can be achieved either by enforcing tailored software (such as TOM-Skype [22]) or by permitting only unencrypted VoIP communication (as provided, e.g., by Empathy [2] or Ekiga [1]). This creates the special situation that also the VoIP clients are considered under control of the attacker, not only the communication and network between the clients. As a consequence, not only encoded packages can be observed by the attacker, but also all input to the VoIP client at the sender and receiver sides. This attacker model is strong but realistic in the considered cencorship context. The censors are further able to observe and eavesdrop on the traffic and to run any statistical analyses.

### 3.2 Design Goals

We consider the following two main design goals for our censorship circumvention system:

1. **Undetectability and Secrecy of Information:** With an attacker observing the communication between two end users as well as their audio input to the VoIP clients, the presence of a hiding scheme should be *inconspicuous* in case of eavesdropping. This restriction is satisfied when the modulation of the audio input leads to an unobtrusive background noise and statistical analyses are not able to distinguish the hidden communication from regular communication. This must also apply for the packet flows and the statistical characteristics of transmitted signals. Even if the activity and characteristics of the circumvention scheme were compromised, the transmitted information should not be accessible for an attacker/censor.
2. **Technical and Organizational Robustness:** Data completeness is important for preserving the content of information. That is, the proposed system must provide a sufficient amount of accuracy within the process of reconstructing the hidden information as well as a *reliable transmission system*. For being reliable, it must recognize and identify fragments of information that were lost or transmitted out of order (e.g., due to unreliable delivery of UDP packets). This might be an issue during forced congestion at networking devices along the transmission path, as well as with packet loss due to temporarily decreased network health.

Furthermore, a censorship circumvention system can only be considered reasonable, if it is publicly accessible for users within a censored area. Therefore it must provide a sufficient amount of flexibility for being combined with an arbitrary VoIP client. We thus present general performance parameters for arbitrary carrier systems and show the system's feasibility in a real-world scenario exemplary with a Skype prototype. Because of its wide deployment and use of the up-to-date audio codec Silk this represents a relevant use case.

## 4 Scheme Description

In the following we provide a high-level overview (Figure 2) of our approach and explain its building blocks. We focus on the central modulation scheme that hides and recovers the secret information. We refer to this as the *general* system setup. Furthermore, we introduce an optional acknowledgment module for increasing the system's transmission reliability.

### 4.1 System Components

The circumvention system provides a modulation module that organizes the hiding and recovering of secret information based on an audio input signal and shared noise sequences. It is attached on top of the VoIP client and utilizes its existing transmission infrastructure. The system components are as follows:

- *Modulation:* The modulation module performs the dynamic hiding and recovering of secret data based on an audio input signal and offers two alternate hiding techniques. Furthermore,

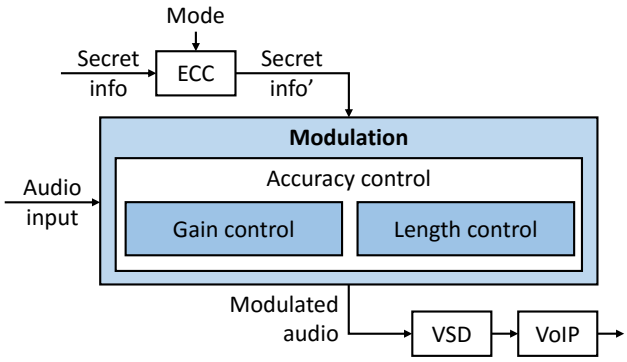


Figure 2: High-level system overview (sender side). The ECC, VSD, and VoIP modules are modular and can be replaced according to the requirements of a deployment scenario.

it provides an accuracy control module that is capable of adapting the modulation process to the audio signal dynamics; this defines the system's accuracy and robustness.

- *Error correction:* To overcome bit errors, the ECC module encodes the secret information input. We use Reed-Solomon (RS) codes for correcting burst errors in full blocks of bits and Golay codes (GC) for correcting single bit errors.
- *Acknowledgment scheme:* The acknowledgment scheme is an optional pre-modulation feature to the general system setup. It enhances the secret information by a frame structure that enables detecting and recovering the loss of information during the transmission process.
- *External components:* SkypeLine can be deployed with an arbitrary VoIP client. A virtual sound device (VSD) acts as an interface between the modulation module and the audio input of the VoIP client.

SkypeLine is run on the application layer and does not interfere with any VoIP protocols. Therefore no alterations to the carrier system are required.

As error correction and external components are reusing existing modules, we proceed to describe only the newly developed modulation and acknowledgement schemes in detail.

## 4.2 Modulation Scheme

The modulation scheme can be operated with two alternate hiding procedures, namely a binary and an m-ary modulation algorithm. We will present a formal specification of both algorithms in the following.

### 4.2.1 Mode A: Binary Modulation

In modulation mode A, each single bit of the secret message is spread over the shared noise sequence to be added to the audio signal. Along with this additional background noise the audio signal can be transmitted by the VoIP client while carrying the hidden bits of the secret message.

The modulation function for a message bit $s_x$ and the noise sequence $\vec{n}$ of length $\vartheta$, applied at the **sender**, is defined as $\vec{m}_x := s_x \cdot \vec{n}'_x + \vec{a}_x$ or:

$$\begin{pmatrix} m_{i,x} \\ \vdots \\ m_{j,x} \end{pmatrix} := s_x \cdot \begin{pmatrix} n'_{i,x} \\ \vdots \\ n'_{j,x} \end{pmatrix} + \begin{pmatrix} a_{i,x} \\ \vdots \\ a_{j,x} \end{pmatrix}, \tag{3}$$

where $\vec{m}_x$ is the modulated signal vector of an interval $[i, j]$ for a secret intermediary bit $s_x \in S = \{-1, 1\}$, $\vec{n}'_x$ is the scaled noise string of length $\vartheta = j - i + 1$, and $\vec{a}_x$ is the audio input sequence of the current interval.

The dynamic scaling of the noise sequences is performed by multiplying each PN sequence with an interval's scaling factor $SF$:

$$\vec{n}'_x := \begin{pmatrix} n_i \\ \vdots \\ n_j \end{pmatrix} \cdot SF_x. \tag{4}$$

This step is performed in the gain control module, where the $SF_x$ is a function of the spectral energy density $E_x$ of the current signal interval $\vec{a}_x$ and a factor $b$. This base factor is defined according to the required noise level (see Eq. 15 for details) and can be optimized via simulation. It represents the volume basis of the background noise for an adjustment to the dynamics of the audio signal:

$$SF_x := E_x \cdot b, \tag{5}$$

$$E_x = \sqrt{\sum_{n=i}^{j} a_{n,x}^2} \tag{6}$$

The accuracy of the demodulation depends on the dynamics of the audio input, as high-energy intervals interfere with the modulated noise sequences. Active gain control improves the distinguishability of the noise and the audio signal within the demodulation step, leading to a higher overall precision of the modulation scheme as well as a stable signal-to-noise ratio (SNR) throughout the entire signal. It adjusts the static volume of the noise sequences to the varying energy of the audio signal. As the noise volume is adapted interval-wise, the gain in precision can be implemented without an expense on the session's SNR. The SNR between a signal $x(n)$ and noise $y(n)$ is defined as follows:

$$SNR_{dB} = \frac{P_x}{P_y} = 10 \cdot \log_{10}\left(\frac{P_x}{P_y}\right) \tag{7}$$

$$P_x = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{-N}^{N} |x(n)|^2. \tag{8}$$

Similar to the gain control step, the length control is performed by increasing the *length* of an interval's noise sequence $[i, j]$. Longer noise sequences increase the probability of a successful demodulation at the expense of decreased throughput rates. The length control is performed in case the interval energy $E_x$ exceeds a threshold $t$ which is defined according to the expected performance of the modulation process: while higher thresholds allow for bandwidth-efficient modulation with less doubling of sequence lengths, a lower threshold provides robust hiding and recovery of information. The length $dim(\vec{n})$ for a current noise sequence $\vec{n}$ is determined as follows:

$$dim(\vec{n}) = \begin{cases} \vartheta, & E_x < t \\ 2\vartheta, & E_x \geq t \end{cases} \tag{9}$$

For a deterministic length control the basic string length $\vartheta$ and threshold $t$ are part of the shared secret. That is, the modulation offsets for all bits of the hidden message can be computed based on the received audio signal.

On the **receiver** side, the hidden information is reconstructed from the received signal $\hat{M}$. For this purpose, both communication partners share the secret information about the noise generator that is used for the current session. The shared secret can consist of specific generation parameters (seed to a PRNG) or a specific key representing the applied noise sequence. Based on this information, the demodulation of an information bit $\hat{M}_x$ is defined as follows:

$$\hat{M}_x := \begin{cases} 1, & \left(\sum_{k=i}^{j} \hat{m}_{k,x} \cdot \vec{n}\right) > 0 \\ 0, & \left(\sum_{k=i}^{j} \hat{m}_{k,x} \cdot \vec{n}\right) \leq 0 \end{cases}, \tag{10}$$

where $\hat{M}_x$ is the recovered information bit, $\hat{m}_{n,x}$ is the current sequence of the received audio signal, and $\vec{n}$ is the shared secret.

We enhance the hiding capabilities of mode A by repeating the modulation step of Eq. 3 for one audio input signal. This results in parallel layers of modulation, that is, one audio input is enhanced by multiple layers of background noise that each carry bits of the secret message. To perform the parallel mode A, the modulated output of the first iteration is processed again with a second noise string and bit vector of the secret information. This adds another layer of background noise on top of the already processed signal, while the use of different noise sequences still allows the demodulation of both layers. For extracting the different layers of noise, the demodulation process is performed in the reverse order of the modulation operations.

In comparison to other DSSS-based approaches, which only provide one layer of modulation, this allows for an increased throughput. The number of parallel modulations on a single audio input is limited by the minimum required accuracy and SNR (see Section **??**). This is due to the fact that additional layers of modulation increase the amount of background noise which results in a decreased SNR.

#### 4.2.2 Mode B: m-ary Modulation

In order to increase the achieved throughput further, we propose a second modulation technique. In m-ary modulation, we use a code book to encode words of the secret message instead of single bits. It is a square Hadamard matrix, where each column represents one word of the code book. Formally, all secret information is encoded in $\Xi = [1, 2, 4t], t \in \mathbb{N}$ orthogonal noise sequences, where $\Xi$ is the dimension of the Hadamard matrix providing the noise sequences. In such a matrix all columns are mutually orthogonal and entries are either 1 or $-1$. In order to encode information we do not spread single bits of information over a noise sequence as in mode A but let each column of the matrix represent a word of length $m$ bits. A matrix of dimension $\Xi$ can encode $m = \lfloor \log_2(\Xi) \rfloor$ bits of information per sequence as there are $2^m = \Xi$ possible column vectors to choose from and therefore words in a code book $W$:

$$W = \begin{cases} \vec{n}_1 \\ \vec{n}_2 \\ \vdots \\ \vec{n}_\xi \end{cases}, \tag{11}$$

where each word represents $m$-bits of the binary alphabet, $\xi$ is the number of possible noise sequences that can be applied to an audio sequence and $\vec{n}_\mu$ is the $\mu$-th word in $W$. Likewise, the Hadamard matrix has $\xi$ columns and $\vec{n}_\mu$ are columns in the matrix.

This technique enables modulating multiple bits through a single noise string instead of representing only one bit as in modulation mode A. This leads to a slightly adapted modulation function at the **sender**, $\vec{m}_\mu := \vec{n}_\mu + \vec{a}$:

$$\begin{pmatrix} m_{i,\mu} \\ \vdots \\ m_{j,\mu} \end{pmatrix} := \begin{pmatrix} n_{i,\mu} \\ \vdots \\ n_{j,\mu} \end{pmatrix} + \begin{pmatrix} a_i \\ \vdots \\ a_j \end{pmatrix}, \tag{12}$$

where $\vec{m}_\mu$ represents the modulated output of the signal interval $[i, j]$ and the $\mu$-th word in $W$.

At the **receiver** the secret information is extracted by correlating each possible word of the shared code book $W$ with intervals of the received signal. The accuracy of the correlation step is affected positively by the pairwise orthogonality of all noise sequences. Based on this characteristic, the m-ary demodulation function applied at the **receiver** side is given as follows:

$$\hat{\xi} := \arg \max_{\mu=1\ldots\xi} (\vec{m}_\mu \star \vec{n}_\mu) \tag{13}$$

$$(f \star g)[n] \stackrel{\text{def}}{=} \sum_{m=1}^{j} f^*[m] g[m+n], \tag{14}$$

where the cross-correlation $\star$ of each column vector $\vec{n}_\mu$ of the noise matrix is computed for the signal and noise sequence vectors $\vec{m}, \vec{n}$. The maximum result in the set of cross-correlations leads to the index $\hat{\xi}$ of the noise string within the matrix of orthogonal strings. Similarly to modulation mode A, the noise sequence volume is adapted dynamically by a gain control module. However, a manipulation of sequence lengths via the length control function would interfere with the set of available orthogonal codes of fixed length; hence there is no length control applied in modulation mode B.

## 4.3 Acknowledgment Scheme

The described modulation modes affect only the audio capabilities and cannot overcome burst errors that arise from packet loss along the connectionless transmission path. Therefore, we propose an optional acknowledgment scheme that enables to detect and recover packet-wise loss of information and extends the original system architecture (see Figure 3). Different to conventional protocols for reliable transmission, our acknowledgment scheme operates on the application layer and therefore does not require any alterations to the carrier protocol that otherwise would result in an attack vector.

The Public Information Hideway (PIH) controller organizes the acknowledgment and retransmission of secret information that is lost during the transmission process. For this purpose the information is converted into a defined frame pattern. These acknowledgment frames are used as input for the modulation module, similar to the original system architecture. All following steps remain the same.

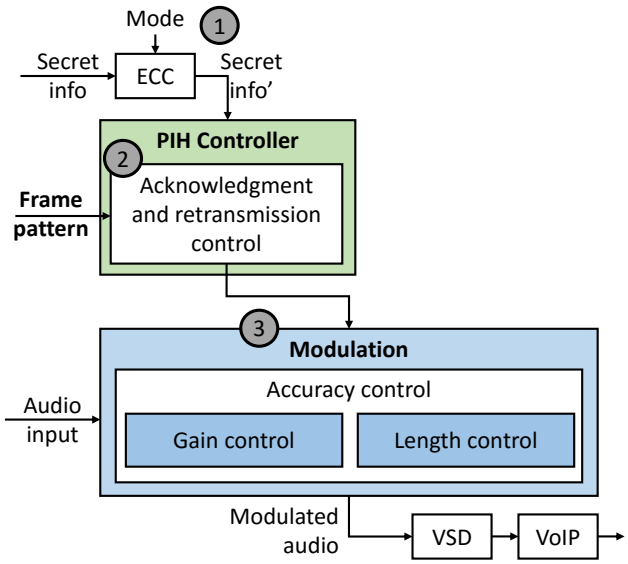


Figure 3: High-level system overview (sender side) including the PIH controller. Note that the PIH is optional and the ECC, modulation, VSD, and VoIP modules remain as described in Figure 2. Numbered circles highlight the steps of the *acknowledgment protocol* (see Section 4.3.2).

### 4.3.1 Acknowledgment Frame

One drawback of the VoIP carrier system is that the variable length encoding of an audio codec fragments the continuous audio stream in packets of differing lengths. By that, the sender and receiver are unable to track the portions of information spread over the transmitted packets. Given this lack of control, we extend the basic modulation scheme by a frame pattern that enables the receiver to reconstruct the original information stream of the sender. For this purpose the original

structure of secret information is transformed from a bit vector into an acknowledgment frame $AF$. Each acknowledgement frame contains at least[2] a sequence number $q$ and a fixed amount of secret information data: AF := ($q$|data), where the frame fields have fixed lengths of $k$ and $l$ bytes, respectively, and $|AF| = k + l$. The field lengths can be optimized with respect to a specific deployment scenario, especially the fragmentation behavior of the VoIP client's audio codec. We use the AF structure to organize the stream of secret information, therefore it does not interfere with the fragmentation of the audio signal that is performed in the VoIP client.

#### 4.3.2 Acknowledgment Protocol

The detection and retransmission of lost information follows the acknowledgment protocol. It initiates the modulation, fragmentation, and sending of secret information at the transmitter as well as the registration of successfully received frames and according acknowledgments at the receiver.

For the sender side the protocol steps are as as follows (see Figure 3 for reference):

1. **Spreading.** In the first step, the secret information is spread bit-wise over the shared noise sequence and is encoded via the ECC module. The output of this module is concatenated to a bit vector ($a$) that sequentially carries all encoded information bits within the noise sequences.
2. **Fragmentation.** The full bit vector ($a$) is fragmented in frames of $l$ bytes length. Each frame is assigned a sequence number in ascending order. Output of the acknowledgment and retransmission control is a new bit vector ($b$) following the predefined frame pattern. It is now a sequence of AFs, where each frame carries $l$ bytes of ($a$).
3. **Modulation.** The bit vector of AFs ($b$) is input for the modulation step that is performed identically to the original modulation process.

At the receiver, each successfully transmitted packet gets accepted by the VoIP client and is reconstructed to an audio signal. In the demodulation module the hidden information is then reconstructed resulting in recovered AF vectors $(b)'$. Based on the sequence numbers of each AF the bit vector $(a)'$ is recomposed and missing sequences can be detected.

For detecting missing frames and initializing the according retransmission the protocol is as follows:

1. **Demodulation.** For recovering the transmitted frame structure the hidden information must be recovered according to the original demodulation process.
2. **Evaluation.** During the ongoing demodulation of the received audio signals, the PIH controller gathers all frames of secret information with their respective sequence numbers. In a periodic evaluation step these sequence numbers are checked for missing frames. All frames that did not arrive since the last evaluation step are considered lost and documented in an acknowledgment message.
3. **Acknowledgment.** The acknowledgment message is sent on the reverse VoIP channel following the exact same hiding process. This event is triggered at the end of each evaluation step; the length of an evaluation period is variable and can be adapted to the robustness requirements of an individual session.
4. **Retransmission.** As soon as the sender receives and recovers an acknowledgment message, it injects the set of missing frames in the ongoing modulation process. To do so, the respective portions of the secret message are framed again with their original sequence number and forwarded to the modulation module. As soon as a retransmission is demodulated successfully, it is removed from the list of missing receptions.

[2]The structure can be extended further by delimiter fields. These fields represent the start and end of an $AF$ through a predefined bit pattern. A further extension would increase the overall robustness at expense of the goodput.

# 5 Experimental Evaluation

In this section we present results from three experimental studies. First, we analyze the performance of the two modulation modes A and B based on results of an extensive simulation study. To prove the feasibility of the proposed system in a real-world scenario, we furthermore present results of a protoype implementation—exemplary for Skype. The optional PIH extension is later discussed based on a network simulation model that enables us to study the acknowledgment capabilities in different packet loss situations of the underlying network.

## 5.1 Simulation Study

We analyze the performance of our modulation scheme based on an extensive simulation study. By this we can optimize the system parameters and show the best case performance for a general system setup.

### 5.1.1 Preliminaries

We define the following preliminaries as basic setup for the simulation study.

*Selection of noise code.* The correlation characteristics of the signal change with the selection of the (pseudo) noise generator. Based on modulation mode A, a White Gaussian Noise (WGN) generator was used that produces sequences similar to white noise resulting in a more natural hearing experience (confirmed in the audibility study below).

To encode m-ary information in modulation mode B, the applied pseudo noise must provide suitable correlation characteristics, as multiple sequences must be distinguishable in the reconstruction step. Therefore, a Hadamard matrix, as introduced in Section 4.2.2, was used for the modulation scheme. It provides a defined number of orthogonal codes that can be distinguished in the reconstruction process.

*Error correction encoding.* In modulation mode A, especially single bit errors occur through the modulation and demodulation process. This was expected, since the VoIP client's audio codec disrupts the reconstruction of single noise sequences, typically leading to single bit errors. We tested Hamming codes and binary Golay codes for mode A. Among these two, the binary Golay code performed better. Even though both codes provide similar data rates, the Golay code reaches the accuracy criterion at a lower sequence length and thus higher data rate.

For the m-ary modulation scheme B, mostly burst errors occur through the modulation process. This is due to the fact that each noise sequence represents a defined number of secret information bits. Thus, a failed reconstruction at the receiver leads to a block error for all bits represented by the sequence. For modulation B, the Reed-Solomon (RS) code is hence more suited, as it is more capable of correcting such burst errors.

*Audibility study:* Since hiding the secret data should be unobtrusive to an attacker, the noise volume should be small in relation to the underlying signal (voice) stream. This can be achieved by defining a lower bound $SNR_{min}$ between the audio input $P(t)$ and noise signal $\tilde{P(t)}$ that must be exceeded at all times:

$$\frac{\frac{1}{j+i-1}\int_i^j P(t)dt}{\int_i^j \tilde{P}(v)dv} \geq SNR_{min}, \tag{15}$$

where $[i, j]$ is the respective interval and $P, \tilde{P}$ are both signal powers.

For evaluating the obtrusiveness of the induced background noise in the presented approach, we performed two audibility studies with a total of 30 participants. The audio files were obtained from meeting recordings of the AMI corpus [12], whereof 20 segments of 23 s length were presented to each participant of the study. High-quality signals from the close-talking microphones were used, which ensures a high signal-to-noise ratios, phases of silence, and the dynamics of realistic conversations making the data as challenging as possible for information hiding.

Results showed that no single bound could be derived for representing one explicit threshold, as the detection accuracy of all participants highly depends on a person's knowledge of the presence

of information hiding as well as expectations about its acoustic form. Additionally, the acoustic characteristics of a pseudo noise generator also affect the perception threshold. A more detailed analysis of the test setup as well as results are given in Section 6.

*Accuracy.* SkypeLine is capable of transmitting any type of data that can be transformed to a binary representation. While media files, e.g., image or video files, are robust to minor errors and can still provide an acceptable quality, more sensitive information like encryption keys is destroyed by single errors. To present the lower bound of SkypeLine's performance in a most restrictive scenario, we assume the task to be the transmission of sensitive information and define an accuracy bound $> 99\%$. That is, only results with a demodulation success of more than 99% are accepted throughout all experiments presented in the following.

### 5.1.2 Experiments

In the following we present the results of our simulation study. This includes the experimental variation of modulation schemes with the objective to further improve the achievable hiding rates. Ensuing from these improvements we summarize the final optimal results that represent the system's best-case performance for a general setup.

*Effects of Gain Control.* The gain control algorithm increases the overall robustness of the modulation and demodulation procedures and limits the conspicuousness of added background noise. In the following we briefly explain the effects of this algorithm.

Based on the amplitude modulation, as described in Section 4.2, noise sequences are added sample-wise to the amplitude of the audio input signal. Thereby, the distances between the audio input and noise sequence amplitudes can differ (see Figure 4). As a consequence, the probability of successfully distinguishing the hidden information from the carrier signal differs from sample to sample, which results in an unstable demodulation process.

To compensate differing distances, the gain control algorithm analyzes the energy of each interval and adapts the scaling factor accordingly. By that, we can apply a minimum distance between the carrier signal and added noise sequences. This increases the robustness of the demodulation step and overall allows for a dynamic modulation of secret information.

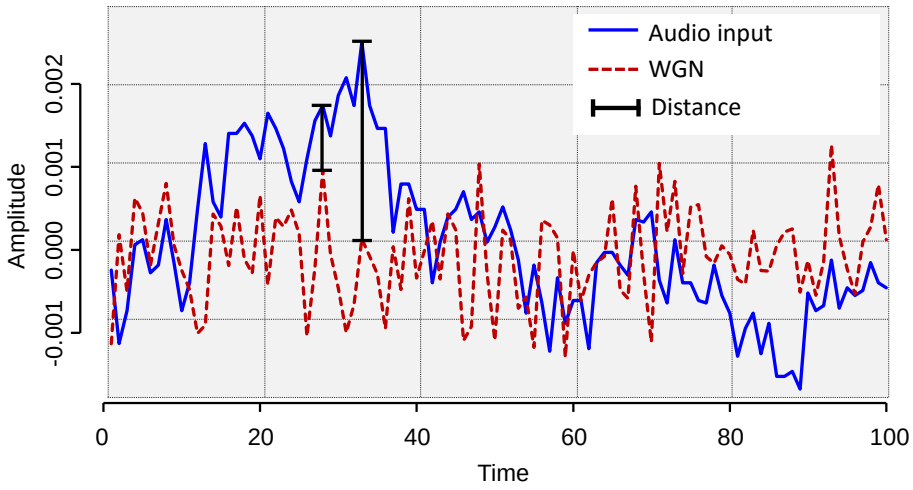


Figure 4: Distances between samples in the audio input and the WGN sequence differ due to the characteristics of both signals.

Besides increased robustness, the dynamic modulation also leads to a harmonized SNR of the background noise (see Figure 5). Hence the overall conspicuousness can be controlled by defining a minimum SNR bound that is guaranteed by the gain control algorithm.

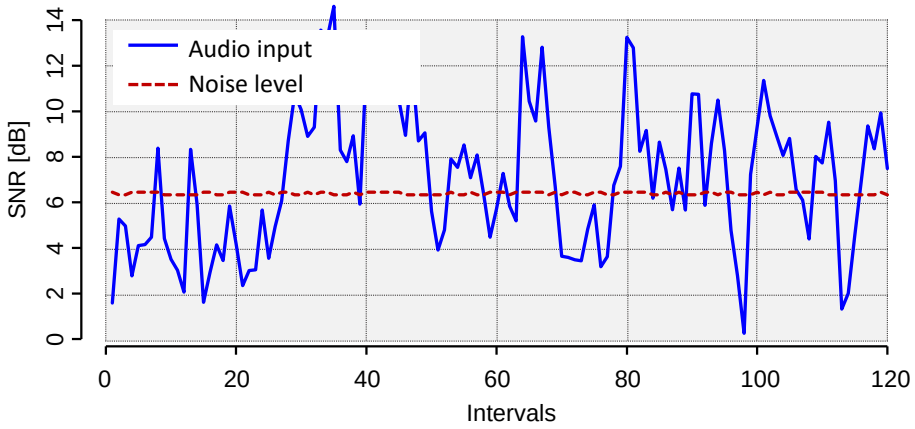


Figure 5: Gain control scales background noise as function of an interval's energy resulting in an overall harmonized SNR. This allows for a robust demodulation at the receiver and leads to a constant background noise level.

*Variations of Modulation Schemes:* The performance of the basic modulation technique, introduced in Section 4.2, can be increased by extending the modulation procedure. In modulation mode A this is realized by repeating the modulation step which results in multiple parallel modulations in one transmitted signal. As a consequence, the overall throughput is increased at the expense of a slightly decreased SNR.

We tested the maximum possible number of parallel modulations with respect to the assumed accuracy requirements that must be met for all layers of hidden information: The reconstruction process can only be considered successful in case all parallel demodulation steps reach 99% correctness.

Up to 4 parallel modulations were tested in the simulation study. This number of modulation layers could be reconstructed with the required accuracy and applies for a general system setup without consideration of any audio codec. Figure 6 shows the summarized results for all simulation runs. The results include all combinations of parameters that were tested for two, three, and four modulation layers and represent the system's overall performance in terms of achievable throughput rates. We will discuss the best case parameter setups in detail as *optimal results* in the following.

The simulation setup with an intermediate OPUS [3] encoding only allows for double modulation up to approx. 72 bps at an SNR of 16 dB, as further iterations could not satisfy the accuracy bound. Even though the optimizations of OPUS result in a comprehensively constrained performance, the results still represent a proof of concept even for compressed audio. While the proposed scheme focuses on flexible deployment with arbitrary VoIP clients, a specialized implementation that adapts to a specific audio codec would allow for increased throughput rates.

*Optimal Results.* The general results in Figure 6 aggregate the results of all possible parameter combinations that were tested in the simulation setup. A box represents 50 % of all results, the upper and lower horizontal bars are the upper and lower 25 % of results. The horizontal bar within the box is the median of all values. Points represent single outliers. This responds to the full performance spectrum of SkypeLine and includes the lower and upper bound of performance that can be achieved by parallel modulation in mode A. For a concluding analysis we now focus on the maximum possible throughput of SkypeLine and therefore repeat the measurements with the strongest parameter sets of the prior experiments. The results show the theoretical upper bound of performance for SkypeLine in mode A and B, whereas the real-world performance is proven based on the prototype implementation in Section 5.2.

With mode A, a repetitive modulation of multiple parallel noise sequences provides the most performant system setup. In this setup we tested noise sequence lengths in range $|N_{i..j}| \in$

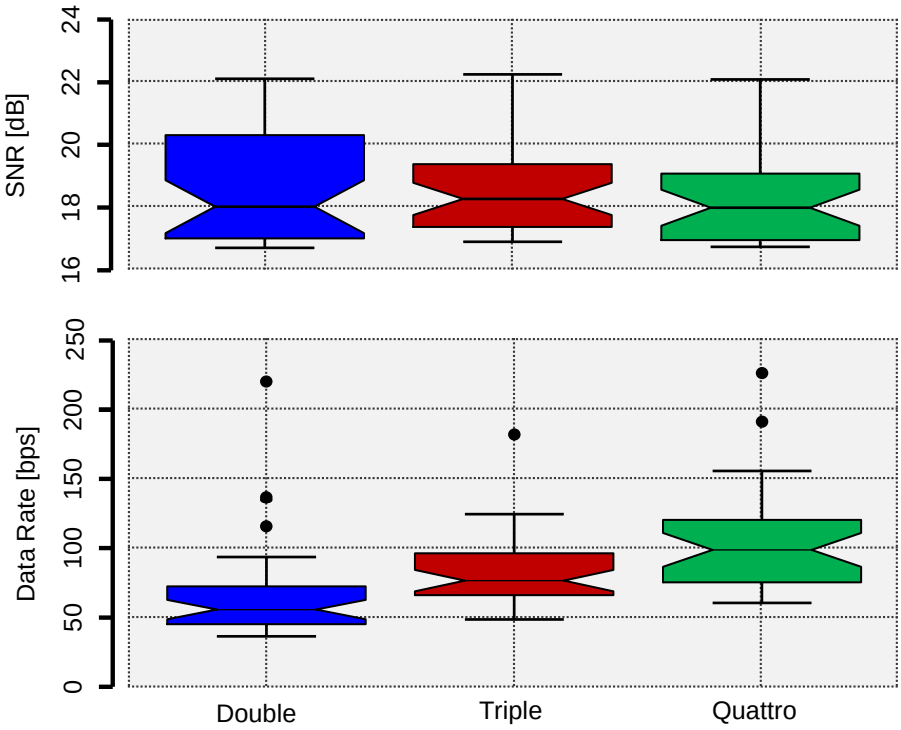


Figure 6: Aggregate SNR and data rate for 720 simulation iterations in mode A. The results fulfill the accuracy constraint of at least 99% successfully recovered secret message bits and therefore are acceptable for all data types.

$[900, 2800]$, base factors in range $b \in [0.02, 0.2]$, and thresholds in range $t \in [0.02, 0.05]$. We conducted multiple repetitions for each parameter combination and filtered all results that did not fulfill the design goals of undetectability and secrecy of information (see Section 3.2). As shown in Table 1, the median and mean data rate of all repetitions summarize the overall performance of the modulation scheme, while the peak data rates represent the best single sample of all simulation runs.

In comparison to recent work in the field of spread-spectrum based steganography approaches, the available data rate is increased significantly, as shown in Tab. 2: We compare the mean data rates of Tab. 1 with the work presented in the work of Takahashi et al. [37] that achieves 20.5 bps for a Frequency Hopping Spread Spectrum (FHSS) system and the work presented by Nutzinger et al. [32] that provides a data rate of 12.5 bps based on a phase-coding hiding technique. Both approaches selected for comparison apply the presented hiding techniques to auditive media and were tested in a context similar to SkypeLine.

We derived the above results from parameter sets that achieve the design goals defined in

| **Mode** | **Peak** | **Median** | **Mean** |
|---|---|---|---|
| Mode A: Single | 36 | 21.98 | 23.54 |
| Mode A: Double | 218 | 55.54 | 67.90 |
| Mode A: Triple | 180 | 76.47 | 85.52 |
| Mode A: Quattro | 224 | 98.63 | 106.61 |
| Mode B | 2,407 | 2,407 | 2,407 |

Table 1: Achievable data rate [bps] of DSSS-modulation scheme without audio codec.

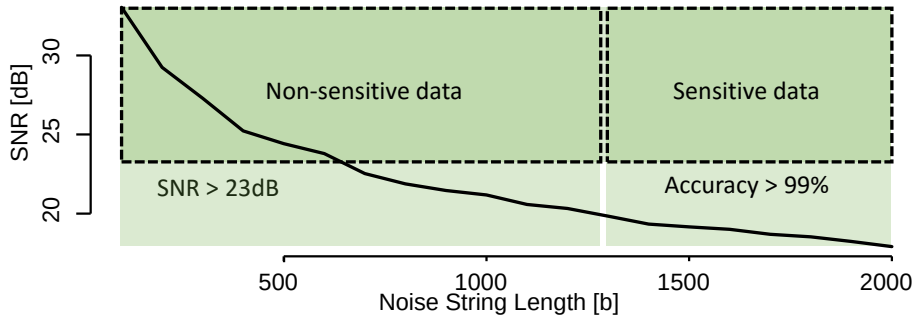


Figure 7: Satisfiability of design goals for increasing noise string lengths. For longer noise strings the phases of higher background noise are prolonged at sequences that have a higher gain factor.

Section 3.2. The maximum attainable SNR is limited by the noise sequence length (see Fig. 7), which is induced by the combination of gain and length control. An increased robustness in the modulation procedure also leads to a decreased SNR and vice versa.

With the alternative mode B, the maximum achievable data rate is increased further. We tested Hadamard matrix dimensions in range of $\Xi \in \{2^2, 2^3, ..., 2^{10}\}$. Compared to the peak data rate of four parallel modulations in mode A, the maximum performance with Hadamard sequences of 128 bit length in mode B provides a performance increase of approx. 1.075 %. This gain in throughput can be achieved at the expense of an increased computational complexity in the demodulation step. Nevertheless, there are no real-time requirements for the recovery of information, thus, this additional computational effort is negligible for most scenarios.

Overall, the achievable throughput can be increased significantly by the proposed modulation techniques. This applies for the extended basic modulation model A, where up to four parallel modulations allow for an improvement of up to factor eight. Furthermore, the novel modulation mode B introduces the possibility of hiding codewords of multiple bits length in single spreading sequences. This represents a major increment in comparison to existing SS-based hiding techniques, as confirmed by the comparison to previous approaches.

In our experiments we strictly followed the SNR and accuracy bound defined before. The presented results therefore represent a general setup in the most restrictive set of constraints. To find a suitable tradeoff between accuracy, background noise, and throughput that fits an individual deployment scenario, these constraints can be reduced and SkypeLine can be extended further, e.g., by additional redundancy.

## 5.2 Prototype Implementation

To complement the theoretical results of the simulation studies, we present exemplary prototype results for the Skype VoIP client that prove the feasibility of our scheme under realistic circumstances. Different to the results of the simulation study, where we focused on the theoretical performance spectrum of the novel modulation techniques, the goal of the prototype experiments is limited to prove the real-world applicability of SkypeLine. Therefore we focus on non-parallel

| **Mode** | **[37]** | **[32]** |
|---|---|---|
| Mode A: Single | 117% | 188% |
| Mode A: Triple | 380% | 684% |
| Mode A: Quattro | 473% | 853% |
| **Mode B** | **12,035%** | **19,256%** |

Table 2: Relative performance improvement in comparison to recent DSSS-approaches.

| Noise | ECC | Throughput |
|---|---|---|
| GC | None | 56 bps |
| GC | Golay | 32 bps |
| WGN | None | 64 bps |
| WGN | Hamming | 40 bps |

Table 3: Selection of average throughput for prototype setup with Skype. Presented results provide at least 99 % accuracy at an SNR of at least 23 dB and were selected from the full set of measurement combinations.

modulation mode A to give a general proof-of-concept in a real-world deployment.

The experimental setup consists of a C++ prototype implementation of the modulation modules, Skype clients at the transmitter and the receiver, and utilizes VBCable [38] as an interface between Skype and the modulation modules. For all results we require the same minimum reconstruction accuracy of 99 % and SNR of 23 dB to reflect the former simulation results. Furthermore, we use close-talk speech recordings [12] of 32 min length as the hiding medium.

Different from the simulation setup, our prototype implementation must be robust to the effects of Skype's speech quality optimization and the performance of its audio codec Silk. As the performance of SkypeLine relies on the individual characteristics of different noise types, we conducted the measurements for two types of sequences, namely White Gaussian Noise (WGN) and Gold codes (GC).

Table 3 summarizes the average achievable throughput for different noise types and ECCs. We conducted these results from measurements with multiple parameter sets and selected the best performing setups that satisfy the SNR and accuracy constraints. Based on these optimal parameters, we repeated 10 transmissions for each setup. Results show that WGN without any ECC applied allows for a throughput of up to 64 bps. This is due to the fact that the correction capabilities of any ECC are induced at the expense of an additional overhead to the secret message. In a robust parameter setup this overhead is not required, as the number of errors is negligible and therefore allows for a more efficient hiding process. Even though the results in Table 3 only allow for very specific use cases, this nevertheless proves the real world applicability of SkypeLine under strict accuracy and SNR constraints.

## 5.3 Acknowledgment Scheme

In order to investigate the scheme's performance under realistic network conditions, we performed a simulation in the network simulation framework OMNeT++ [33]. Given a network model of a transmitting and receiving node as well as intermediate networking devices, the effects of variable network parameters were measured.

**Preliminaries.** The length of an acknowledgment frame (AF) affects the tradeoff between performance in terms of minimum overhead and robustness in terms of maximum success ratio in the recovery of frames. To this end we optimized the data field length of an AF towards a setup that provides the best achievable robustness. This requirement is fulfilled for a data field length of 300 bytes, that is, in the following simulations each AF carries 300 bytes of the secret message along with the respective sequence number. Note that the AF length does not interfere with expected packet lengths at the transmission channel, as the acknowledgment scheme performs on application layer, as depicted in Figure 3.

**Simulation Study.** We analyzed the acknowledgment scheme in scenarios of increasing packet loss rates. The transmission protocols and traffic patterns were selected to reflect the VoIP carrier in the simulation model. In particular, this includes UDP as protocol and setting the packet lengths and packet transmission rates according to the Wireshark captures of real Skype traffic (see Table 4 for details).

As shown in Figure 8, the capability of retransmitting lost frames leads to an extension of

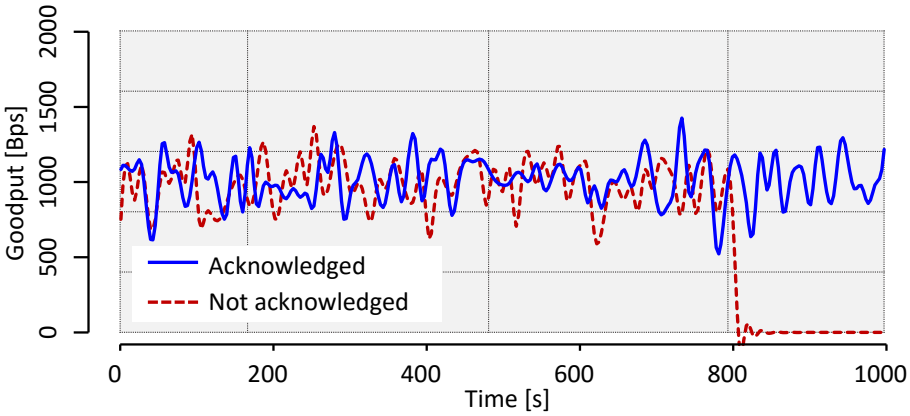


Figure 8: Comparison of goodput for acknowledged and unacknowledged mode at a high packet loss rate of approx. 56%.

the overall transmission process, while the overall amount of received data is increased by 20%. With an unacknowledged performance of the proposed scheme, all packets lost due to a decreased network health reduce the information quality of received data. When transmitting loss-critical information, this can result in a complete loss of quality. However, ongoing retransmissions of lost data as occurring with an acknowledged mode lead to an increased data reliability even in scenarios of high packet loss.

# 6 Security Analysis

Our security analysis consists of three parts. We investigate (1) the resistance to statistical mismatches of the communication patterns, (2) unobstrusiveness of the hidden communication as investigated in a hearing study, and (3) the statistical indistinguishability of the audio signals.

## 6.1 System Mismatches

According to Geddes et al. [13], the vulnerability of a censorship circumvention scheme can be analyzed with respect to three major types of mismatches.

- **Architectural mismatches:** The circumvention scheme modifies the architecture of the cover protocol or differs from it, e.g., a client server architecture deployed over a peer-to-peer host system. This allows an attacker to identify and block nodes possibly involved in hidden communication.
- **Channel mismatches:** Different requirements in the reliability of transmissions can lead to data loss at the receiving host. This may, e.g., be an issue that occurs when transmitting loss-sensible data over a connection-less transmission protocol. Packet loss can be induced by an attacker by targeted packet dropping or duplication and forced congestion at networking devices.
- **Content mismatches:** When the expected content differs from the embedded content of the covert channel, differences in traffic patterns are noticeable by the censor. The detection capability is increased by variable length encoding resulting in packet lengths characteristic for different speech and language sets.

We assure that our proposed circumvention scheme provides resistance against attacks based on the three types above.

As the modulation of hidden information is performed on real speech data, the data transmission can be performed by the hosting VoIP system without alteration of the underlying system architecture. Hence, neither additions to the cover protocol nor changes in the existing protocol

| | Mean | | SD | |
|---|---|---|---|---|
| **Mode** | **IAT** | **PL** | **IAT** | **PL** |
| No modulation | 0.019 | 130.34 | 0.001 | 13.87 |
| SNR inaudible | 0.019 | 126.98 | 0.001 | 12.72 |
| SNR audible | 0.019 | 122.1 | 0.001 | 12.52 |

Table 4: Packet lengths (PL) and inter-arrival times (IAT) of captures for 5 audio sequences of unmodulated data [12], 150 minutes in total.

structure are necessary for the implementation of our proposed scheme. For interceding the transmission of hidden information, the censor must block the transmission or reception directly at the connection ends. This implies that the censor simultaneously blocks architecture elements that are crucial for the carrier system's performance.

Even though the proposed scheme is based on a connection-less carrier system, the acknowledgment of transmissions can help to prevent data loss by packet loss, e.g., due to single drops of packets or phases with high packet loss rates in the underlying network. The number and identity of lost packets can be recovered due to the sequence numbering in received packets. Retransmissions of missing packets would then be invoked by sending acknowledgment messages to the transmitting host—these messages would be embedded in the existing VoIP stream in the same hidden way as the original information.

Content mismatches are detectable when a statistical analysis of the transmissions' traffic patterns reveals conspicuous differences between original and altered data of the circumvention scheme. To analyze the resistance of the proposed modulation scheme against statistical analysis, we used Wireshark [4] to capture genuine traffic and traffic with hidden information. We then compared them in terms of packet lengths (PL) in bytes and inter-arrival (IAT) times in seconds. In the experimental setup, we sent non-modulated and modulated audio data of 10 minutes length between two Skype nodes (two virtual hosts on one physical machine), while capturing the incoming traffic at the receiving host. The experiments used five different audio signals that were sent in genuine format or contained hidden information in an either inaudible (SNR $> 25\,dB$) or audible (SNR $< 25\,dB$) way. Our results are shown in Table 4. There were no statistically significant differences in the packet inter-arrival times. While slight differences in the mean values of the packet lengths can be noticed, the standard deviations disguise this effect for specific transmissions and do thus not reveal conspicuous differences between the original and modulated data, in particular for the targeted inaudible modulations. A statistical analysis performed solely on the packet stream will therefore not be successful for detecting content mismatches between the carrier system's original data and the adapted data. The results of a more detailed statistical analysis of the time and transform domain signal are given in Section 6.3.

## 6.2 Unobtrusiveness

In case of eavesdropping, the activity of SkypeLine must not be detected through the new acoustic characteristics of the modulated signal. That is, the recording of a VoIP conversation that carries secret information must be similar to a standard conversation. We tested the subjective conspicuousness of altered and unaltered signals within a second audibility study and compute the PESQ (perceptual evaluation of speech quality) index for a standardized reference value in comparison to common VoIP speech quality indexes.

While the initial audibility study focused on determining a modulation threshold to scale the background noise, the second study with another 15 participants was conducted with the objective to quantify the obtrusiveness of modulated audio signals. The participants were presented modulated and unmodulated audio files that already provide a natural background noise in the original signal; the test set was derived from movie snippets containing conversations of two or more characters. The setup included a set of 6 original and 24 modulated *.wav* files with different scaling factors in range of 19 dB $<$ SNR 40 dB. All participants received an identical set of *.wav*

| Mode | Mean | SD |
|---|---|---|
| **Original** | 3.22 | 7.57 |
| **Modulated** | 3.28 | 7.56 |

Table 5: Mean and standard deviation of signal amplitude after FFT.

files, while the playback order was generated randomly. Additionally, all participants were provided with knowledge about the modulation procedure and its acoustic form. This includes the fact that the set of audio samples consisted of modulated and original pairs of sound files played randomly with a 50 % portion of modulated files.

An SNR of 34 dB was best performing in our test: In this case, only 35 % of participants suspected the modulation activity, which results in a success rate (unnoticed hidden communication) of 65 %. The success rate decreased to approx. 57 % for larger SNRs, indicating unstable detection rates. This fact is confirmed by 42 % of detections where participants supposed modulations in the original audio files (false positives), while 38 % of all modulations could not be detected at all (false negatives). Based on these results, the probability of specifically detecting the circumvention activity of the proposed system by eavesdropping can be considered low: within the audibility study, only a limited number of participants were able to *reliably* detect modulations. Through all tests, no participant was able to determine the complete set of original audio files.

Additionally, we computed a reference PESQ score to compare the subjective results to the standardized test methodology, where higher PESQ scores indicate a better audio quality: Computed for the file set of the above audibility study, the mean PESQ for all files is 4.1879. Most VoIP codecs degrade the PESQ-scores to values below 4.0 [5, 7], so that our modifications do not degrade the overall call quality and, hence, do not induce alterations to the expected acoustic characteristics.

## 6.3 Statistics of Audio Signals

Another technique for detecting activity of the proposed circumvention system is a statistical analysis of the transmitted audio signals. While the audibility study led to unstable detection rates for eavesdropped signals, a statistical analysis can reveal inaudible characteristics of a modulated signal that differ from standard signals of VoIP conversations. To study the resistance of our scheme to statistical analyses, two investigations were performed: a comparison of the frequency domain after a fast Fourier transform (FFT) and a $\chi^2$ homogeneity test.

### 6.3.1 Frequency domain analysis

The absolute FFT amplitudes of the original and the modulated signals were compared with respect to their mean and standard deviation. The test set consisted of 140 `.wav` signal pairs of 30 s length with 70 original and 70 modulated signals with 7 scaling factors resulting in a noise level within 19 dB $<$ SNR $<$ 40 dB. In total, 490 minutes of close-talk original and modulated signals, derived from the AMI corpus [12], were considered in the FFT analysis.

In Table 5, the mean and standard deviation of frequencies are summarized for the original and modulated signals. While the mean amplitude of modulated signals differs by approx. 1.85 % from the original signal, the standard deviation only differs by approx. 0.12 %, leading to minimal differences in the frequency domain.

### 6.3.2 $\chi^2$ test

In a second statistical analysis, we applied the $\chi^2$ homogeneity test for 70 audio signal pairs of 30 s length from the AMI corpus [12], modulated with 7 scaling factors resulting in noise loads within 19 dB $<$ SNR $<$ 40 dB according to a method proposed by Provos and Honeyman [35]. *Homogeneity* is given if a set of random samples can be considered equally distributed. A distribution indicates equality if the histograms of all random samples are equal, which is set as

| **SNR** $[dB]$ | 19 | 22 | 25 | 26 | 29 | 31 | 39 |
|---|---|---|---|---|---|---|---|
| **Prob.** [%] | 42.86 | 61.43 | 77.13 | 77.13 | 80.1 | 82.86 | 82.86 |

Table 6: Results of the $\chi^2$ test showing the probability of an equal distribution for increasing SNR values of the modulation.

the null hypothesis $H_0$ of the test. For deciding whether a signal pair is indistinguishable, the null hypothesis must be accepted within a pre-defined significance level $\alpha$, allowing for frequency discrepancies up to the $\alpha$-threshold.

The homogeneity of a signal pair is given by the $\chi^2$ test value. It is computed for the $k$-category histograms of the time and frequency domain representation of the original and modulated signal:

$$\chi^2 = \sum_{j=1}^{k} \sum_{i=1}^{m} \frac{(n_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{16}$$

$$E_{i,j} = \frac{n_{i\bullet} n_{\bullet j}}{n_k}, \tag{17}$$

where $k$ is the histogram granularity and $m$ is the number of signals in the homogeneity test ($m = 2$ since we test signal pairs). $E_{i,j}$ is the marginal probability of each category and denotes the expected number of occurrences for a category $k_j$ and signal $m_i$. In the matrix representation of the test input, the categories $k_j \in \{k_1, k_2, \cdots, k_l\}$ are column by column and signals $m_i \in \{m_1, m_2\}$ are row by row. Within the matrix, $n_{i,j}$ are the individual matrix values, and $n_{i\bullet}$,$n_{\bullet j}$, and $n_l$ are the column sums, the row sums, and the total number of occurrences, respectively.

The $\chi^2$ test values are added up for all rates in the cells of the matrix leading to the overall test value for a signal pair. If this test value is within the predefined $\alpha$-quantile, $H_0$ is accepted and the signal pair can be considered as *equally distributed*:

$$\chi^2 \leq \chi^2_{(k-1)(m-1);(1-\alpha)}, \tag{18}$$

where $\alpha$ is the significance level of the test. The probability of a modulated signal passing the homogeneity test is defined as the $p$-value of the test pair and derived from the inverse $\chi^2$ cumulative distribution function with $v = k - 1$ degrees of freedom:

$$p = 1 - \int_0^{\chi^2} \frac{t^{(v-2)/2} e^{-t/2}}{2^{v/2}\Gamma(v/2)} dt, \tag{19}$$

where $\Gamma$ is the Euler Gamma function.

Table 6 summarizes the relative success for each set of signal pairs in a time domain analysis, where higher probabilities indicate better indistinguishability and therefore statistical security of SkypeLine. Results show that for an SNR of at least 25 dB the rate for successfully distinguishing modulated from original signals is below 25 %. We emphasize that the presented analysis can only be conducted when an attacker (i.e., the censorship authority) has access to the original version of a transmitted signal. Whenever an unpredictable carrier signal is used (e. g., for live voice transmissions or unpredictable recordings), this analysis cannot be performed since a pairing of the cover medium and the modulated signal is not feasible. Since it is an easy task for the transmitter to use an unpredictable and fresh audio recording, this can be considered realistic.

# 7 Discussion

In the following, we emphasize the circumvention potential of SkypeLine and discuss the impact of audio codecs.

### 7.1 Deployment Scenario

Steganography in the context of censorship circumvention must provide undetectability as its principal security objective. The proposed system and its performance should therefore only be compared to steganographic approaches that guarantee security features similar to, or better, than those given with our DSSS system.

Use cases for the proposed system are restricted by the assumed attacker model that implies direct blacklisting and eavesdropping of communication channels. In this context, low-bandwidth services, e.g., messaging, key establishment [34], or the exchange of Tor bridges, represent beneficial deployment scenarios. Even though alternate approaches, such as TRIST [9] or Facade [19], provide higher throughput, they cannot be applied to VoIP as a legitimate carrier system. Even though a deployment with alternate carrier technologies is possible, VoIP is most reasonable in context of censorship circumvention. In this context, SkypeLine allows for an improved goodput for DSSS-based systems while focusing on protection against strong attackers instead of rich-data deployment scenarios.

### 7.2 Impact of Audio Compression

Even though our system can be applied with arbitrary VoIP clients, the overall throughput highly depends on the compression functions of the VoIP client's audio codec as well as the audio optimization capabilities. Modern codecs, such as Silk [20] and Opus [3], use vector quantization (VQ) and therefore compress signal data by estimating approximate vector representations that are optimized with respect to high speech quality and noise reduction. For the modulation technique of SkypeLine, this affects the overall reliability, as the quantization process filters the noise sequences carrying the hidden information. The impact of a VQ-based codec can be reduced by selecting highly-configurable VoIP clients allowing for an adaption of the compression and channel parameters.

The Opus version of Silk is optimized for high speech quality and noise shaping, whereas Celt is applied in high-bandwidth channels providing a compression that is more suitable for our modulation technique. The use of Celt can be forced by configuring a covert channel of at least 20 kbps. This leads to a more reliable reconstruction of hidden information and was applied in our simulation study. Further adaptions are possible, e. g., on the complexity parameters of the compression mode.

## 8 Related Work

There is a large body of previous work that investigates censorship circumvention systems. In the following, we relate existing work with our approach and especially focus on steganography and watermarking schemes.

To overcome the restrictions of Internet censorship, several systems and circumvention schemes have been proposed; Mazurczyk [25] provides a comprehensive survey of the existing literature. Such schemes can be based on inter-packet delays or losses, or on modifying the packet payload. Among the latter, the Least Significant Bit (LSB) technique and the spread-spectrum technique are common examples.

With the LSB technique [6, 8, 10, 24], negligible bits within the carrier signal are replaced by bits of secret information. As this step can be applied to each signal sample, it provides high throughput rates. However, each LSB scheme must be adapted to the carrier system's specific codec and therefore is limited in portability. Asad et al. [6] present an LSB scheme that reduces the probability of detection by randomizing the bit pattern used for hiding purposes. Liu et al. [24] propose the Least Significant Digit (LSD) method that overcomes capacity restrictions of low bit rate speech by fully utilizing frame bits.

Regarding spread-spectrum based steganographic techniques, FHSS- (Frequency Hopping) and DSSS-based approaches have been utilized for hiding secret information in audio signals. Nutzinger et al. [31] present a hybrid steganography approach that combines FHSS and DSSS. This hybrid

technique is based on a DSSS modulation but, different from our scheme, allows for a variation of the carrier frequency through FHSS. Takahashi & Lee [37] compare the performance and robustness of covert channels based on low-bit coding (LBC), echo data hiding, FHSS, and DSSS. Metrics for this evaluation are the applicable bandwidth, feasibility in respect of computation times, and robustness to signal processing. Nutzinger & Wurzer [32] utilize a phase-coding mechanism as an alternative covert channel. Different from our work, the phase coding of secret information is performed after a fast Fourier transform of the audio input, leading to a manipulation of the frequency representation of the signal.

Zielinska et al. [43] present a DSSS-modulated steganography approach for the IEEE802.15.4 Wireless Personal Area Network (WPAN) standard with the goal to minimize the probability of a covert channel disclosure while being robust to random errors. The proposed system allows for increased data rates of up to $250\,kb/s$ at the expense of a decreased SNR for the modulated output. Even though this throughput rate allows for a performant utilization of the covert channel, the proposed system does not provide undetectability in case of an eavesdropping censor, as the SNR undercuts the defined threshold of an unobtrusive steganography system.

Hamdaqa et al. [14]present a method for providing increased reliability without weakening the steganography system. It is based on the LACK approach [27] and depends on intentionally delayed multimedia packets that are discarded by the receiver. Nevertheless, both implementations depend on nearly loss-free transmissions for critical packets which cannot be provided for standard VoIP traffic. The results of [30,36,41] also refer to a VoIP carrier and provide different advantages regarding reliability and efficiency, but are based on unrealistic drop rates.

# 9 Conclusion

In this paper, we have presented a DSSS-based steganography approach that hides information in a standard VoIP communication system. The hiding is performed by modulating information bits onto the voice carrier signal using pseudo-random noise sequences; different from prior approaches we use orthogonal noise sequences and repeat the spreading operation. The resulting signals are transmitted by a VoIP client. The data rates achieved with our technique substantially exceed existing DSSS approaches.

In our evaluation, we have focused on the robustness of the demodulation and inaudibility. With an acknowledgment scheme, the data completeness can be improved significantly even for limited network resources while increasing the robustness against manipulations of the packet flow by a censor. Based on two simulation studies, we have shown that the parametrized modulation and acknowledgment schemes can be adapted to the characteristics of a specific deployment scenario, providing a more performant censorship circumvention system.

# 10 Acknowledgments

The research presented in this paper was supported by the DFG Research Training Group GRK 1817/1.

# References

[1] Ekiga: Softphone, video conferencing, and instant messenger. `http://www.ekiga.org`, 2014.

[2] Empathy messaging program. `http://live.gnome.org/Empathy`, 2014.

[3] Opus codec. `http://www.opus-codec.org/`, 2014.

[4] Wireshark. `http://www.wireshark.org/`, 2014.

[5] Broadcom codec comparison. `http://www.broadcom.com/support/broadvoice/codec_comparison.php`, 2015.

[6] M. Asad, J. Gilani, and A. Khalid. An enhanced least significant bit modification technique for audio steganography. In *Computer Networks and Information Technology (ICCNIT), 2011 International Conference on*, 2011.

[7] H. Assem, D. Malone, J. Dunne, and P. O'Sullivan. Monitoring voip call quality using improved simplified e-model. In *Computing, Networking and Communications (ICNC), 2013 International Conference on*, 2013.

[8] R. Cogranne and F. Retraint. An Asymptotically Uniformly Most Powerful Test for LSB Matching Detection. *Information Forensics and Security, IEEE Transactions on*, 2013.

[9] Christopher Connolly, Patrick Lincoln, Ian Mason, and Vinod Yegneswaran. Trist: Circumventing censorship with transcoding-resistant image steganography. In *4th USENIX Workshop on Free and Open Communications on the Internet (FOCI 14)*, 2014.

[10] Tomáš Denemark and Jessica Fridrich. Detection of content adaptive lsb matching: a game theory approach. 2014.

[11] Roger Dingledine and Nick Mathewson. Design of a blocking-resistant anonymity system. Technical report, The Tor Project, 2006.

[12] Carletta et al. The AMI Meeting Corpus: A Pre-announcement. In *Second International Conference on Machine Learning for Multimodal Interaction*, 2006.

[13] John Geddes, Max Schuchard, and Nicholas Hopper. Cover your acks: Pitfalls of covert channel censorship circumvention. 2013.

[14] M. Hamdaqa and L. Tahvildari. Relack: A reliable voip steganography approach. In *Secure Software Integration and Reliability Improvement (SSIRI), 2011 Fifth International Conference on*, 2011.

[15] Amir Houmansadr, Chad Brubaker, and Vitaly Shmatikov. The Parrot is Dead: Observing Unobservable Network Communications. 2013.

[16] Amir Houmansadr, Giang T. K. Nguyen, Matthew Caesar, and Nikita Borisov. Cirripede: Circumvention Infrastructure using Router Redirection with Plausible Deniability. In *Computer and Communications Security*. ACM, 2011.

[17] Amir Houmansadr, Thomas Riedl, Nikita Borisov, and Andrew Singer. I want my voice to be heard: IP over Voice-over-IP for unobservable censorship circumvention. In *Network and Distributed System Security*, 2013.

[18] Luca Invernizzi, Christopher Kruegel, and Giovanni Vigna. Message In A Bottle: Sailing Past Censorship. In *Annual Computer Security Applications Conference*, 2013.

[19] Ben Jones, Sam Burnett, Nick Feamster, Sean Donovan, Sarthak Grover, Sathya Gunasekaran, and Karim Habak. Facade: High-throughput, deniable censorship circumvention using web search. In *4th USENIX Workshop on Free and Open Communications on the Internet (FOCI 14)*, 2014.

[20] K. Soerensen K. Vos, S. Jensen. RFC Draft: Silk. 2009.

[21] Jeffrey Knockel. Tom-skype research. `http://cs.unm.edu/~jeffk/tom-skype/`, 2013.

[22] Jeffrey Knockel, Jedidiah R. Crandall, and Jared Saia. Three Researchers, Five Conjectures: An Empirical Analysis of TOM-Skype Censorship and Surveillance, 2011.

[23] Zhen Ling, Xinwen Fu, Wei Yu, Junzhou Luo, and Ming Yang. Extensive Analysis and Large-Scale Empirical Evaluation of Tor Bridge Discovery. In *INFOCOM*, 2012.

[24] Jin Liu, Ke Zhou, and Hui Tian. Least-significant-digit steganography in low bitrate speech. In *Communications (ICC), 2012 IEEE International Conference on*, 2012.

[25] Wojciech Mazurczyk. Voip steganography and its detection - a survey. *ACM Comput. Surv.*, 46(2):20:1–20:21, December 2013.

[26] Wojciech Mazurczyk, Maciej Karaś, Krzysztof Szczypiorski, and Artur Janicki. Youskyde: information hiding for skype video traffic. *Multimedia Tools and Applications*, 2015.

[27] Wojciech Mazurczyk and Józef Lubacz. LACK - a VoIP Steganographic Method. *CoRR*,

2008.

[28] Wojciech Mazurczyk, Paweł Szaga, and Krzysztof Szczypiorski. Using transcoding for hidden communication in ip telephony. *Multimedia Tools and Applications*, 70(3):2139–2165, 2012.

[29] Hooman Mohajeri Moghaddam, Baiyu Li, Mohammad Derakhshani, and Ian Goldberg. SkypeMorph: Protocol Obfuscation for Tor Bridges. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, 2012.

[30] H. Neal and H. ElAarag. A packet loss tolerant algorithm for information hiding in voice over IP. In *Southeastcon, 2012 Proceedings of IEEE*, 2012.

[31] M. Nutzinger, C. Fabian, and M. Marschalek. Secure hybrid spread spectrum system for steganography in auditive media. In *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2010 Sixth International Conference on*, 2010.

[32] M. Nutzinger and J. Wurzer. A novel phase coding technique for steganography in auditive media. In *Sixth International Conference on Availability, Reliability and Security (ARES)*, 2011.

[33] OMNeT++ Community. OMNeT++ Network Simulation Framework. `http://www.omnetpp.org/`, 2014.

[34] Abdelkader H. Ouda and Mahmoud R. El-Sakka. A step towards practical steganography systems. In *Proceedings of the Second International Conference on Image Analysis and Recognition*, ICIAR'05, 2005.

[35] N. Provos and P. Honeyman. Detecting steganographic content on the internet. `http://www.citi.umich.edu/techreports/reports/citi-tr-01-11.pdf`, 2001.

[36] R. Roselinkiruba and R. Balakirshnan. Secure steganography in audio using inactive frames of voip streams. In *Information Communication Technologies (ICT), 2013 IEEE Conference on*, 2013.

[37] Takehiro Takahashi and Wenke Lee. An Assessment of VoIP Covert Channel Threats. In *Third International Conference on Security and Privacy in Communications Networks and the Workshops. SecureComm 2007*, 2007.

[38] `http://vb-audio.pagesperso-orange.fr/Cable/`. Vbcable, 2015.

[39] Qiyan Wang, Xun Gong, Giang T. K. Nguyen, Amir Houmansadr, and Nikita Borisov. CensorSpoofer: Asymmetric Communication using IP Spoofing for Censorship-Resistant Web Browsing. In *Computer and Communications Security*. ACM, 2012.

[40] Zachary Weinberg, Jeffrey Wang, Vinod Yegneswaran, Linda Briesemeister, Steven Cheung, Frank Wang, and Dan Boneh. StegoTorus: A Camouflage Proxy for the Tor Anonymity System. In *Computer and Communications Security*. ACM, 2012.

[41] Erchi Xu, Bo Liu, Liyang Xu, Ziling Wei, Baokang Zhao, and Jinshu Su. Adaptive voip steganography for information hiding within network audio streams. In *Network-Based Information Systems (NBiS), 2011 14th International Conference on*, pages 612–617, 2011.

[42] Zinan Yang, HongZhi Liu, Hongzhi Liu, and Xiaogang Lv. The applied research on internet of things engineering surveillance's records management. In *Software Engineering and Service Science (ICSESS), 2012 IEEE 3rd International Conference on*, pages 689–693, June 2012.

[43] E. Zielinska and K. Szczypiorski. Direct sequence spread spectrum steganographic scheme for IEEE 802.15.4. In *Multimedia Information Networking and Security (MINES), 2011 Third International Conference on*, 2011.