Guesswork and Entropy as Security Measures for Selective Encryption

Reine Lundin
Guesswork and Entropy as Security Measures for Selective Encryption

Reine Lundin
Guesswork and Entropy as Security Measures for Selective Encryption

Reine Lundin

DISSERTATION

Karlstad University Studies | 2012:36

ISSN 1403-8099

ISBN 978-91-7063-443-7

© The author

Distribution:
Karlstad University
Faculty of Economic Sciences, Communication and IT
Computer Science
SE-651 88 Karlstad, Sweden
+46 54 700 10 00

Print: Universitetstryckeriet, Karlstad 2012

WWW.KAU.SE
Guesswork and Entropy as Security Measures for Selective Encryption

REINE LUNDIN
Department of Computer Science
Karlstad University
Sweden

Abstract

More and more effort is being spent on security improvements in today’s computer environments, with the aim to achieve an appropriate level of security. However, it might be necessary for small computing devices to reduce the computational cost imposed by security in order to gain reasonable performance and/or to decrease energy consumption. Selective encryption that provides confidentiality by encrypting only chosen parts of the information can be used to accomplish this. Previous work on selective encryption has mainly focused on how to reduce the computational cost while still making the information perceptually secure, not on how computationally secure the information is.

Despite the efforts made and due to the complex nature of computer security, good quantitative assessment methods for computer security are still lacking. Inventing new ways to measure security in general and selective encryption in particular are therefore necessary in order to better understand, assess and improve the security of computer environments. Two proposed probabilistic quantitative security measures are entropy and guesswork. Entropy gives the average number of guesses in an optimal binary search attack, and guesswork gives the average number of guesses in an optimal linear search attack. In information theory, a considerable amount of research has been carried out on entropy and on entropy-based metrics. However, the same does not hold for guesswork.

In this thesis, we evaluate performance improvement when using the proposed generic selective encryption scheme. We also examine the confidentiality strength of selectively encrypted information by using and adopting entropy and guesswork. Moreover, since guesswork has been less theoretically investigated than entropy, we extend guesswork in several ways and investigate some of its behaviors.

Keywords: computer security, security metrics, selective encryption, confidentiality, entropy, guesswork.
Acknowledgments

First and foremost I would like to express my gratitude to my main supervisor Stefan Lindskog and co-supervisor Simone Fischer-Hübner. They have encouraged, guided, and supported me in my research studies during the whole time.

I also would like to thank my other co-authors of publications appended in this thesis, namely Anna Brunstrom, and Thijs Holleboom, for the many valuable and stimulating discussions, and all colleagues at the Computer Science department for providing such a nice working place. Additionally, I would like to thank Anders Johansson for the innumerable discussions about nothing and everything ranging from theoretical considerations to everyday perplexities. Finally, I dedicate this thesis to my two children Martin and Rebecka Lundin for providing so much joy.

This work was financially supported by the Knowledge Foundation of Sweden with TietoEnator and Ericsson as industrial partners, county administrative board of Värmland, and the European Regional Development Fund (ERDF) through the Compare Business Innovation Centre phase 2 (C-BIC 2) project, due to which I am grateful.

In memorial of my father

Karlstad, August, 2012

Reine Lundin
List of Appended Papers

This thesis is based on the work presented in the following eight papers. References to the papers will in the introductory summary be made using the roman numbers associated with the papers.


Some of the papers have been subjected to minor editorial changes.

Comments on My Participation

In paper I, I have contributed with ideas, discussions and some written material concerning selective encryption. However, most of the written material and performance evaluations is accomplished by Stefan Lindskog. For papers II–VIII, I am responsible for most of the written material and ideas while Stefan Lindskog, Simone Fischer-Hübner, Anna Brunstrom and Thijs Holleboom have proofread and commented on ideas.
Other Papers

Apart from the papers appended to the thesis, I have also authored or co-authored the following papers:


## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>i</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iii</td>
</tr>
<tr>
<td>List of Appended Papers</td>
<td>v</td>
</tr>
<tr>
<td>Comments on My Participation</td>
<td>vi</td>
</tr>
<tr>
<td>Other Papers</td>
<td>vii</td>
</tr>
<tr>
<td>Introductory Summary</td>
<td>1</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>3</td>
</tr>
<tr>
<td>2 Related Work</td>
<td>5</td>
</tr>
<tr>
<td>3 The Generic Selective Encryption Scheme</td>
<td>6</td>
</tr>
<tr>
<td>4 Security Measures</td>
<td>8</td>
</tr>
<tr>
<td>4.1 A Note on Measurements</td>
<td>8</td>
</tr>
<tr>
<td>4.2 Entropy</td>
<td>10</td>
</tr>
<tr>
<td>4.3 Guesswork</td>
<td>11</td>
</tr>
<tr>
<td>4.4 The Relationship between Entropy and Guesswork</td>
<td>13</td>
</tr>
<tr>
<td>5 Confidentiality Strength of Selective Encryption</td>
<td>14</td>
</tr>
<tr>
<td>5.1 Neighborhoods</td>
<td>14</td>
</tr>
<tr>
<td>5.2 Guessing</td>
<td>16</td>
</tr>
<tr>
<td>6 Research Questions</td>
<td>20</td>
</tr>
<tr>
<td>7 Research Methodology</td>
<td>21</td>
</tr>
<tr>
<td>8 Main Contributions</td>
<td>22</td>
</tr>
<tr>
<td>9 Summary of Papers</td>
<td>23</td>
</tr>
<tr>
<td>10 Concluding Remarks and Future Work</td>
<td>27</td>
</tr>
<tr>
<td>Paper I: Middleware Support for Tunable Encryption</td>
<td>31</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>33</td>
</tr>
<tr>
<td>2 Related Work</td>
<td>35</td>
</tr>
<tr>
<td>3 Middleware Design</td>
<td>36</td>
</tr>
<tr>
<td>3.1 Architectural Overview</td>
<td>36</td>
</tr>
<tr>
<td>3.2 Block-based Selective Encryption</td>
<td>37</td>
</tr>
</tbody>
</table>
Introductory Summary
1 Introduction

Computer security is an important branch of computer science and an increasing part of overall society, with its purpose to secure entities from being tampered with in an unauthorized manner. The best known way of defining computer security is to divide it into the three categories confidentiality, integrity and availability. Confidentiality is the prevention of unauthorized disclosure of information, integrity is the prevention of unauthorized modification of information, and availability is the prevention of unauthorized withholding of information or resources. They are collectively known as the “CIA”. Another way of defining computer security is to divide it into the categories of prevention, detection, response and recovery, based on where in the line of defense the attacks are resisted [15]. Prevention is when direct resistance against attacks is provided, detection is when signatures of ongoing attacks are searched for in the environment, response is when different countermeasure activities are triggered based on the detected attack and, finally, recovery is when the computer environment is restored from the attack. Other definitions and categorizations of security exist as well, see for example [11, 33].

When protecting security items, a computational cost is imposed on the computer environment. This extra computational overhead might negatively affect performance, e.g., the response time, availability and throughput. For small computing devices with restricted resources, such as mobile phones or sensor network devices, security mechanisms can put a significant extra burden on the performance and energy consumption of the device. Furthermore, security items should be secured according to the adequate security principle [25]. This principle states that the items should only be protected to a degree consistent with their value. Hence, if the value of an item changes over time, so should the protection level.

The concept of selective encryption may be used to reduce computational cost and fulfill the principle of adequate security. Selective encryption is a concept that reduces computational cost when it provides confidentiality by only encrypting chosen parts of the information. Previous work on selective encryption has mainly been aimed towards multimedia applications with short information lifetime, such as TV or radio broadcast of events, in order to reduce the computational cost and/or energy consumption while still making the information perceptually secure to a certain protection level. The perception protection utilizes the fact that different parts of the information have different impacts on our perception senses, i.e., eyes and ears. However, the confidentiality strength in the sense of computational security has only been briefly mentioned or rudimentarily
analyzed. Furthermore, by altering the distribution of encrypted parts, selective encryption can also be used to trade confidentiality against computational cost.

In this thesis, we investigate the performance improvement and examine the confidentiality strength of a proposed generic selective encryption scheme. The performance improvement of the scheme is evaluated in paper I and initial results on its confidentiality strength are presented in paper II by adopting the security measure guesswork to selective encryption. In papers VI and VII, an entropy equation of selective encryption is derived and investigated for different orders of languages. In paper VIII the entropy equation is further extended with information neighborhoods to capture information dependencies in several dimensions, and then applied on bitmap images. In order to better understand guesswork, we generalize and explore some behaviors of guesswork in papers III–V. The relation between entropy and guesswork is examined in paper III, the definition of guesswork is extended to joint and conditional guesswork in paper IV and, finally, an investigation of how guesswork changes over time in multi-processor attacks is conducted in paper V.

The notations and mathematical expressions have changed over time in the appended papers, hopefully thanks to increased knowledge. For instance, the information parts in the selective encryption scheme are sometimes referred to as units or blocks, and the entropy expression for selectively encrypted strings has gone through several notation changes and information encapsulations to increase abstraction. This notation evolution process followed to find a better and more exact description of the research topic under consideration was expressed by Einstein as [1]:

“It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.”

The remainder of this introductory summary is organized as follows. Related work is addressed in Section 2, and the generic selective encryption scheme used is presented in Section 3 together with a short discussion of its performance. Entropy and guesswork are defined in Section 4, and the confidentiality strength of selective encryption is discussed in Section 5. The research questions addressed in this thesis are given in Section 6, the research methodology is presented in Section 7, and the main contributions are stated in Section 8. A summary of the papers included is given in Section 9 and, finally, Section 10 concludes the introductory summary and discusses future work.
2 Related Work

The concept of selective encryption was first and independently introduced by Spanos and Maples [32], Li et al. [16], and Meyer and Gadegast [24] in 1995 and 1996. The purpose was to reduce the amount of encrypted MPEG data in a video sequence by only encrypting a subset of the video frames while still providing an acceptable level of perceptive security. Selective encryption has also been used for H.264/AVC video streams [30] and in a wireless video camera [12] to save energy and processing time.

For uncompressed bitmap images, selective bitplane encryption was investigated in [27, 35] for the purpose of saving energy and computational cost in mobile environments while still protecting image perception. Selective encryption has also been studied for JPEG2000 images [4, 23]. Furthermore, a perception-based selective encryption scheme for telephone data compressed with the ITU-T G.729 8 kb/s speech encoding standard was presented in [29]. To decide which parts of the bit stream from the compression algorithm to encrypt, the authors systematically corrupted a given bit and then measured the corresponding perceptual impact. Selective encryption for the G.729 speech encoding standard has also been studied in [34].

[19] presented an “m-out-of-n” information-independent selective encryption model. The model lets the user or application decide which parts of the information to encrypt by dividing the information into \( n \) equally sized parts. The first \( m \) parts are then encrypted with a strong encryption algorithm and the following \( n - m \) parts with a weaker and faster encryption algorithm. This pattern is repeated for the whole message. A content independent or generic selective encryption service was also presented in [17]. In the service, the user or application decides which parts of the information to encrypt by dividing the information into equally sized parts and then using a bit mask to indicate which parts should be encrypted. The remaining parts are unencrypted.

The confidentiality strength of selectively encrypted information has only been briefly mentioned or rudimentarily analyzed before; see for instance [27]. Moreover, good quantitative assessment methods for computer security are despite numerous efforts still lacking [13, 14, 18, 36]. This is mainly due to the harsh nature of computer security, since attacker environments are difficult to model. To better understand, assess and improve the security of computer environments new ways of measuring security are therefore needed. Two proposed, and in this thesis used, probabilistic quantitative security measures are entropy [31] and guesswork [22, 26]. These two measures are further described in Section 4.
3 The Generic Selective Encryption Scheme

The generic selective encryption scheme used throughout the thesis consists of three entities: the information, $I$, to selectively encrypt, the bit vector, $b$, controlling which parts of $I$ to encrypt, and the selectively encrypted information, $E(I)$. In the scheme, $I$ is divided into the $n$ information parts $I_i$. Hence,

$$I = \bigl\{ I_i \bigr\}_{i=0}^{n-1}$$

where $|$ denotes the binary concatenate operator. A part $I_i$ is then encrypted if $b_i \mod |b| = 1$ and left unencrypted if $b_i \mod |b| = 0$. The modulus operator is used if the number of information parts is larger than the size of the bit vector. From this, the selectively encrypted message is constructed as

$$E(I) = \bigl\{ \begin{array}{ll} I_i & \text{if } b_i \mod |b| = 0 \\ E(I_i) & \text{if } b_i \mod |b| = 1 \end{array} \bigr\}_{i=0}^{n-1}$$

Fig. 1 illustrates the generic selective encryption scheme. Moreover, the size and fraction of the encrypted parts in $E(I)$ naturally define the encryption level, according to the following definition.

![Diagram](image-url)
Definition 1. Let $E(I)$ be selectively encrypted information consisting of $n$ parts. The encryption level, $EL$, is then defined as

$$EL = \frac{\sum_{i=0}^{n-1} |I_i| b_i \mod |b|}{|I|}$$

where $|I|$ and $|I_i|$ denote the size of the information and its parts, respectively.

The generic selective encryption scheme was implemented in paper I in C/C++, and performance was evaluated by using two connected PCs, acting as a sender and a receiver. In the experiment $|I| = 10$ MB, $|b| = 64$ bits and the AES algorithm were used with a 128 bit key; hence, $|I_i| = 128$ bits. The mean encryption and transmission time from the experiment are shown in Fig. 2. As a reference, pure AES encryption was also measured. The figure indicates that the computational overhead scales almost linearly with respect to the encryption level. The overhead produced by the selective encryption scheme can be found by comparing it with the measured time for pure AES encryption. As long as the amount of encrypted information parts is less than or equal to 93% in the prototype implementation, the selective encryption scheme produces less overhead than encrypting everything using pure AES.
4 Security Measures

In this section, the two proposed probabilistic quantitative security measures entropy and guesswork are defined. These two measures provide an indirect indication from different guessing strategies of how computationally secure selectively encrypted information is. A brief note on measurements will however first be presented.

4.1 A Note on Measurements

In measurement theory [9], a measurement produces an instant value of an attribute by using a predefined measure process. The values might be divided into qualitative or quantitative values. Qualitative values have a direct realization by means of a natural language description such as small, medium and large, while quantitative values have an indirect realization by means of numbers, such as 5 or 42. For quantitative values to be understandable, a unit must be added, for instance 8 meters or 2 seconds. Repeated measurements of the same measure can be compared to each other by putting the values on a scale. A scale consists of a set of values corresponding to the range of possible values of the measure under consideration. Five major types of scales exist [9]:

– Nominal
– Ordinal
– Interval
– Ratio
– Absolute

The nominal and ordinal scales use qualitative values, while the interval, ratio and absolute scales use quantitative values.

Despite numerous efforts made to quantify computer security, the security measures are mostly qualitative, i.e., based on experience as in the common criteria [6]. Hence, they do not allow for an analytical and exact description of security and are therefore not very easy to use in order to more exactly understand, assess and improve the security of computer environments. The knowledge development from experience in mathematically proving that something holds unconditionally can be seen in the categorization of the strength of cryptographic algorithms [11]. In the categorization, the strength of cryptographic algorithms that aim to provide confidentiality of information is divided into the following three categories:
4. Security Measures

- Empirically secure
- Provably secure
- Unconditionally secure

Fig. 3 illustrates the relationship between the three categories. They are more formally defined as follows.

**Definition 2.** An algorithm is empirically secure if it has by experience over time a broad acceptance in the community that it provides confidentiality of information as expected without any known drawbacks.

Note that empirically secure algorithms could with new insights be broken tomorrow.

**Definition 3.** An algorithm is provably secure if it can be proved that the cost of breaking it exceeds the value or the useful lifetime of the information being protected.

When the computational speed increases and hardware becomes cheaper, the cost of breaking the algorithm decreases. Hence, it might from time to time be desirable to update the algorithm or increase the key length used by the algorithm in order to maintain that the algorithm is still provably secure. Provably secure is sometimes also referred to as computationally secure.¹

¹ To decide whether or not the information is computationally secure, a computational threshold needs to be determined. The value or size of the threshold is not considered in this thesis.
**Definition 4.** An algorithm is unconditionally secure if it cannot be broken regardless of what resources the attacker has.

Examples of unconditionally secure algorithms are the one-time-pad encryption scheme [2] and the Dining Cryptographers Network (DC-net) [5].

Entropy and guesswork can be used to measure the number of guesses needed to break or find a secret. These two measures are therefore employed in this thesis to indirectly measure how computationally secure selectively encrypted information is. The next two subsections will define and discuss these two measures.

### 4.2 Entropy

The entropy $H(X)$ [31] is the classical measure of uncertainty and was originally suggested by Shannon in 1944. He defined the entropy as the average amount of information of a discrete random variable $X$, where $X$ attains values from the sample space $\mathcal{X} = \{x_1, \ldots, x_n\}$ with the probability distribution $p_i = p(X = x_i)$. From this, entropy is defined as follows.

**Definition 5.** The entropy, $H(X)$, of a random variable, $X$, with probability distribution $p_i$ is defined as

$$H(X) = - \sum_i p_i \log_2 p_i \quad (4)$$

Entropy can also be seen as a measure giving the average number of guesses in an optimal binary search attack; see paper III for details.

The entropy can be extended to the joint and conditional entropy [7]. The joint entropy, $H(X_1, X_2)$, gives the entropy of a pair of random variables with the joint probability distribution $p_{ij} = p(X_1 = x_i, X_2 = x_j)$.

**Definition 6.** The joint entropy, $H(X_1, X_2)$, of a pair of random variables $(X_1, X_2)$ with the joint probability distribution $p_{ij}$ is defined as

$$H(X_1, X_2) = - \sum_{i,j} p_{ij} \log_2 p_{ij} \quad (5)$$

The conditional entropy $H(X_2|X_1)$, or equivocation which it is often referred to, gives the remaining entropy of the random variable $X_2$ given the random variable $X_1$ with the conditional probability distribution $p_{j|i} = p(X_2 = x_j|X_1 = x_i)$. 
4. Security Measures

**Definition 7.** The conditional entropy, $H(X_2|X_1)$, of the random variable $X_2$ given the random variable $X_1$ with the conditional probability distribution, $p_{ji}$, is defined as

$$H(X_2|X_1) = \sum_i p_i H(X_2|X_1 = x_i) = -\sum_{i,j} p_{ij} \log_2 p_{ji}$$

(6)

The marginal, joint and conditional entropies are related through the chain rule [7] as follows

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1)$$

(7)

Hence, on average, the number of guesses needed to find the value of $(X_1, X_2)$ is equal to the number of guesses needed to find the value of $X_1$ and then $X_2$ given the value of $X_1$. Thus, the chain rule makes it possible to decompose the guessing problem additively into smaller pieces. Moreover, the different entropies also possess the following inequalities

$$H(X_1|X_2) \leq H(X_1) \leq H(X_1, X_2)$$

(8)

giving that conditioning reduces entropy and joining increases entropy. Joint and conditional entropies might be generalized to $n$ random variables, and so might the chain rule (7) and the inequalities in (8).

4.3 Guesswork

Guesswork [22,26] is a measure that gives the average number of guesses needed to find the value of a random variable $X$ in an optimal brute force attack. A brute force attack is equal to a linear search attack; see paper III for details. When performing such an attack, the attacker is assumed to have complete knowledge of the probability distribution, $p_i$, of $X$. Hence, before the guessing process starts, the attacker arranges $p$ in a non-increasing probability order. Mathematically, this ordering of $p_i$ can be expressed through a permutation, $\sigma(i)$, that maps the index of the largest value to one, the index of the second largest value to two, and so on until the index of the smallest value is mapped to $n$. From this, guesswork is defined as follows.
**Definition 8.** The guesswork, \( W(X) \), of a random variable \( X \) with probability distribution \( p_i \) that is arranged in a non-increasing probability order by the permutation \( \sigma(i) \) is defined as

\[
W(X) = \sum_i \sigma(i)p_i \tag{9}
\]

As shown in paper IV, in a similar way as in entropy, guesswork can be extended to joint and conditional guesswork. The joint guesswork, \( W(X_1, X_2) \), gives the guesswork of a pair of random variables with the joint probability distribution \( p_{ij} \).

**Definition 9.** The joint guesswork, \( W(X_1, X_2) \), of a pair of random variables, \((X_1, X_2)\), with the joint probability distribution \( p_{ij} \) that is ordered in a non-increasing probability order by the permutation \( \pi(i, j) \) is defined as

\[
W(X_1, X_2) = \sum_{i,j} \pi(i, j)p_{ij} \tag{10}
\]

The conditional guesswork, \( W(X_2|X_1) \), gives the remaining guesswork of the random variable \( X_2 \) given the random variable \( X_1 \) with the conditional probability distribution \( p_{j|i} \).

**Definition 10.** The conditional guesswork, \( W(X_2|X_1) \), of the random variable \( X_2 \) given the random variable \( X_1 \) with conditional probability distribution \( p_{j|i} \) that is ordered in a non-increasing probability order by the permutations \( \rho_i(j) \) is defined as

\[
W(X_2|X_1) = \sum_i p_i W(X_2|X_1 = x_i) = \sum_{i,j} \rho_i(j)p_{ij} \tag{11}
\]

Moreover, the marginal, joint and conditional guessworks possess the same type of inequalities as entropy. Hence,

\[
W(X_1|X_2) \leq W(X_1) \leq W(X_1, X_2) \tag{12}
\]

Thus, conditioning reduces guesswork and joining increases guesswork. However, in contrast to entropy, the different guessworks are not related through the same type of chain rule property, which makes it possible to easily decompose the guessing process additively into smaller pieces.
4. Security Measures

4.4 The Relationship between Entropy and Guesswork

The relationship between entropy and guesswork has been under consideration for a time [3, 21, 22]. Considering Definition 8, the last term in the sum is weighted with $n$. However, the second last guess in the guessing process determines the last two values of the random variable. That is, if the answer to the second last question is correct, then the searched value is $x_{n-1}$ and the search finishes. If instead the answer is incorrect, then the searched value must be $x_n$ and the search also finishes. Based on this reasoning, guesswork has been redefined as follows.

**Definition 11.** The guesswork, $W'(X)$, of a random variable $X$ with probability distribution $p_i$ that is arranged in a non-increasing probability order by the permutation $\sigma(i)$ is defined as

$$W'(X) = \sum_i \sigma'(i)p_i$$  \hspace{1cm} (13)

where

$$\sigma'(i) = \begin{cases} \sigma(i) & \text{if } \sigma(i) < n \\ n-1 & \text{if } \sigma(i) = n \end{cases}$$  \hspace{1cm} (14)

As derived in paper III, the redefined guesswork and entropy are related as

$$W'(X) = H(X) + D(X || 2^{-\sigma'(i)})$$  \hspace{1cm} (15)

where $D(X || 2^{-\sigma'(i)})$ is the relative entropy [7]. The relative entropy, or Kullback-Leibler distance to which it is also referred, is always non-negative and zero if and only if $p = q$. Moreover, the relative entropy can be interpreted as a measure of inefficiency since it gives the extra number of bits needed for a code of an arbitrary distribution than for the code of the “true” distribution. Formally it is defined as follows.

**Definition 12.** The relative entropy, $D(X_1 || X_2)$, of a random variable $X_1$ with probability distribution $p_i$ and a random variable $X_2$ with probability distributions $q_i$ is defined as

$$D(X_1 || X_2) = \sum_i p_i \log_2 \left( \frac{p_i}{q_i} \right)$$  \hspace{1cm} (16)
5 Confidentiality Strength of Selective Encryption

This section presents and discusses initial results on the confidentiality strength of selective encryption. Before describing the confidentiality strength, a subsection discussing information neighborhoods is provided.

5.1 Neighborhoods

In [31], contiguous sequences of symbols from a language, called \( n \)-grams, were used to find the probabilities of the symbols in order to approximate texts in the considered language. The approximation was carried out for different \( n \)-grams as follows:

- In the zero order approximation, \( \omega = 0 \), the symbols are independent and uniformly distributed.
- In the first order approximation, \( \omega = 1 \), the symbols are independent and distributed as they are in the language.
- In the second order approximation, \( \omega = 2 \), symbols are dependent on one preceding symbol and distributed as they are in the language.
- In the \( n \) order approximation, \( \omega = n \), the symbols are dependent on \( n - 1 \) preceding symbols and distributed as they are in the language.

The order gives the size of the \( n \)-grams used and thus determines the set of depending symbols in the approximation, referred to as an information neighborhood in paper VIII. In the remainder of this section, symbols will be referred to as points. Moreover, the neighborhoods might also contain succeeding points and have dependencies in several dimensions when the information has a multi-dimensional representation state. Hence, the order concept needs to be generalized to an order vector, \( \omega \), where the elements give the order in the corresponding dimension. For instance, \( \omega = (2, 1) \) means that the order is two in the first dimension and one in the second dimension.

A neighborhood will be denoted \( \mathcal{N}_\omega^i \), where the sub index gives the order and the super index gives the number of preceding points in the corresponding dimension of the neighborhood. The number of succeeding points is then given by \( \omega - i - 1 \). Fig. 4 illustrates all nine basic neighborhoods of the two-dimensional family \( \mathcal{N}_{3,3}(x, y) = \mathcal{N}_3(x) \times \mathcal{N}_3(y) \). A black square represents the point under consideration, \( \textbf{p} = (x, y) \), and the white squares represent the corresponding preceding or succeeding points of the neighborhoods. The \( \mathcal{N}_{3,3}^{2,1}(x, y) \) axis neighborhood is equal to the von Neumann neighborhood of range one [20]
The nine basic neighborhoods of the two-dimensional family $N_{3,3}(x, y) = N_3(x) \times N_3(y)$. A black square represents the considered point, $p = (x, y)$, and the white squares represent the corresponding preceding or succeeding points of the neighborhoods.

Fig. 4: The nine basic neighborhoods of the two-dimensional family $N_{3,3}(x, y) = N_3(x) \times N_3(y)$. A black square represents the considered point, $p = (x, y)$, and the white squares represent the corresponding preceding or succeeding points of the neighborhoods.

when using the $L_1$ metric [10], which is also equal to the concept of 4-connected pixels in computer images.

The basic neighborhoods contain only points located on the axes. However, all points outside the axes, but within a given distance determined by $\omega$, could also be considered to be in the neighborhood. Using the $L_1$ metric, a neighborhood can be constructed by joining the outermost points in the one-dimensional basic neighborhoods, thereby creating a convex polytope acting as the neighborhood. An $n$ polytope is a set or geometric object in $n$ dimensions with flat sides. A 2-polytope is usually referred to as a polygon and a 3-polytope as a polyhedron. Fig. 5 shows the neighborhood for the basic neighborhood $N_{5,5}(x, y)$. Note the points in the neighborhood that are not included in the basic neighborhood. Furthermore, if the axis neighborhoods are symmetrically located around the considered point, the neighborhoods are actually circles in the $L_1$ metric.
When the neighborhoods have the shape of convex polytopes they can mathematically be described by the convex hull [28]. The convex hull, \( \text{Conv}(S) \), of a set \( S \) of points is the smallest convex polytope that contains \( S \). From this, the neighborhood of depending points can be written as

\[
D_{\omega}^1(p) = \text{Conv}(N_{\omega}^1(p)) \setminus p
\]

where \( \setminus \) is the setminus operation. In the next subsection, constructed information neighborhoods will be used to capture information dependencies in several dimensions.

### 5.2 Guessing

Depending on the pre-knowledge that the attacker possesses before a guessing attack, different strategies can be considered to speed up the guessing process. For instance, knowledge about the probability distribution of the sample space will most likely affect the order in which the attacker guesses. By assuming that the encryption algorithm used, e.g., RSA or AES, does not have any known weaknesses, the attacker is more or less forced to perform a brute force attack on
the key space when trying to break the encrypted information. However, when using selective encryption, the unencrypted parts may leak information to the attacker about the encrypted parts owing to information dependencies, thereby decreasing the message space, consisting of all possible messages, which is typically much larger than the key space. In this thesis, the attacker is assumed to perform a partial ciphertext only attack on the message space. That is, we assume that the attacker has complete knowledge of which parts of the message are encrypted and that the probability distribution on the message space is also known to the attacker. It is also assumed that the attacker knows when the correct message has been found in the guessing attack.

When the probability distribution of the sample space is known, the entropy and guesswork can be used to measure the average number of guesses in the guessing attack. The difference between the two measures resides in which questions are allowed and used in the guessing process. Entropy measures the average number of guesses in an optimal binary search attack, where the environment allows questions on sets. Guesswork, on the other hand, measures the average number of guesses in an optimal linear search attack, where the environment only allows questions on singletons. See paper III for a detailed discussion.

One question that now arises is: What are we actually guessing at in the guessing process? Ideally, the message space consists of all meaningful messages matching the length of the encrypted part. However, generating meaningful messages of a specific length can be a cumbersome task. That is why the approximation model described in the previous subsection is used to construct hypothetical meaningful messages. Moreover, since entropy possesses the chain rule property (\ref{eq:chain_rule}), the same result is achieved if the guessing attack is performed using whole messages or singletons that construct the messages. Guesswork, on the other hand, does not possess the same simple mathematical structure as entropy does, and at the time of this writing no chain rule decomposition for guesswork is known. Hence, most of the initial results on the confidentiality strength of selective encryption are derived for entropy.

To model the behavior of selectively encrypted information, random variables are associated to $E(I)$ as follows:

\[
E(I_i) = \begin{cases} 
X_i = I_i & \text{if } b_i = 0 \\
X_i & \text{if } b_i = 1
\end{cases}
\]  

(18)

Thus, the $X_i$:s will be known if the corresponding part is unencrypted and unknown if the corresponding part is encrypted. Using this association and adopting guesswork to selective encryption, it is shown in paper II for zero order...
approximations that

\[
W(E(I)) = \frac{|\chi|^n + 1}{2}
\]  

(19)

where $|\chi|$ is the size of the alphabet in the considered language. Since guesswork does not seem to possess a simple additive structure, higher order approximations for guesswork are still an open research issue.

For entropy, an equation for selective encryption has been derived for higher orders and in several dimensions in papers VI-VIII to

\[
H^1_\omega(E(I)) = \sum_p b(p) \prod_{p' \in R(p)} p(I_{p'} | I_{\mathcal{D}_\omega(p')}) H(X_p | X_{\mathcal{D}_\omega(p)})
\]

\[
= \sum_p b(p) p^1_\omega(X_\mathcal{R}_\omega(p)) H^1_\omega(X_p)
\]  

(20)

The region $\mathcal{R}_\omega^1(p)$ is a connected subset of the information, restricted by jumps of width $\mathcal{D}_\omega(p)$ over unencrypted areas or by the boundary of the information. How to mathematically express the structure of $\mathcal{R}_\omega^1(p)$ is an issue of future research. Moreover, in paper VIII, the entropy equation (20) is applied on a $512 \times 512$ pixel bitmap version of the famous Lena image, depicted in Fig. 6. Moreover, the probability distributions, $p^1_2(I_z)$, of the Lena image are shown in Fig. 7, where $I_z$ denotes bitplane $z$ in the image. No legend is inserted due to the large amount of plots in the graph. However, bits of equal adjacent values tend to cluster, and this property increases for higher bitplanes. The highest probabilities that occur with value one in bitplane eight are $p^1_2(1|110)$ and $p^1_2(0|001)$. 

\[\text{Fig. 6: The famous Lena image.}\]
Finally, the case of only considering encryption of whole bitplanes and only the dependency of one adjacent bitplane, and assuming a steady state of the product in (20), the two entropies $H^{1,1,1}_{2,2,2}(E(I_z)|b_{z-1} = 1)$ and $H^{1,0,1}_{2,2,2}(E(I_z)|b_{z-1} = 1)$ is shown in Fig. 8.
6 Research Questions

The research questions for this thesis are threefold.

1. How much is the performance improved when the generic selective encryption scheme is used?

The main goal and purpose of selective encryption is to reduce the computational cost when providing confidentiality by only encrypting chosen parts of the information. This research question is addressed in paper I.

2. How much is the confidentiality strength of the information affected when the generic selective encryption scheme is used?

Previous work on selective encryption has chiefly focused on how to reduce the computational cost while still making the information perceptually secure. However, the confidentiality strength in the sense of computationally secure has only briefly been mentioned or rudimentarily analyzed. This challenging question is addressed in papers II, VI and VII. Initial results on the confidentiality strength of selective encryption by using guesswork are presented in paper II, and an equation for the entropy of selectively encrypted strings is derived in papers VI-VII. Moreover, paper VIII constructs and discusses information neighborhoods in order to make it possible in future work to generalize the entropy equation for higher orders and dimensions.

3. How does guesswork relate to entropy and how can guesswork be generalized in order to create a theory for guesswork?

Security measures are needed to investigate how computationally secure selective encryption is. Entropy and guesswork are two measures that are considered to measure the confidentiality strength of selective encryption. Until today, guesswork has been less theoretically investigated than to entropy. Therefore, in order to better understand guesswork, papers III-V investigate some behaviors of guesswork. The relationship between entropy and guesswork is investigated in paper III, and the definition of guesswork is extended to joint and conditional guesswork in paper VI, where it is also proved that the joining of random variables increases guesswork while conditioning of random variables reduces guesswork. Finally, an investigation of changes in guesswork over time in multi-processor attacks is conducted in paper V in order to understand how the average number of guesses changes during the guessing process.
7 Research Methodology

Research can be seen as the process of asking and answering questions in an organized way in order to produce new knowledge. Hence, before exploring a research question, a research method that structures the research process must be decided upon. One structured way of conducting research is given by the following steps [8]:

1. Problem identification
2. Literature study
3. Determine research questions
4. Information gathering
5. Analyzing and interpreting
6. Reporting and evaluating

The first step investigates the overall research focus and, from new knowledge obtained in step two, the third step breaks down the overall research focus to specific research questions or hypotheses. The three specific research questions in this thesis were presented in the previous section, and the overall research focus that goes beyond the scope of this thesis could be stated as: From the viewpoint of quality of service (QoS), could selective encryption be used to trade confidentiality against performance? Moreover, for performance and security analysis, both experimental and analytical approaches are commonly used for the last three steps. In this thesis, the experimental approach is mainly applied in research question one and the analytical approach in research questions two and three. Hence, a major part of this thesis uses the analytical research approach.

In the analytical research approach, mathematical models are used to describe, explain, predict and eventually control properties in the model. Further, by referring to the list of steps above, step four is used to make assumptions, identify unknowns, introduce suitable notations, investigate conditions and, if possible, transform the problem or parts of the problem into a previously known problem. Step five is used to choose a solving strategy and to carry out the actual work of the chosen strategy. The set of strategies and solving skills will of course depend on earlier experience and knowledge. Step six, finally, is used for reflection and generalization in order to investigate whether the derived result is reasonable and extendable. A problem in the analytical approach arises when the system under consideration becomes too complex with many variables. In such situations, assumptions are introduced to simplify the model, thereby increasing the risk of oversimplifying the model and removing many important properties. The strength of modeling with mathematics lies in its exact description, hence
avoiding the possibility of subjective interpretations. Based on this and the fact that the security measures used in this thesis are probabilistically defined, the analytical approach was chosen for research questions two and three.

When validating analytical models or investigating complex systems, the experimental research approach might be used to gain insight into the defined research question. Real world experiments were conducted to investigate research question one. The reasons for using real world experiments were that the selective encryption scheme was already implemented and set up in a controllable environment.

8 Main Contributions

The main contribution of this thesis lies in the fields of selective encryption and security measures. Below follows a summary of the main contributions:

– Selective encryption has earlier been studied for a specific application or content context. In this thesis, we designed a content independent or generic selective encryption scheme as a middleware and investigated its performance gains in paper I. The experimental results show that the selective encryption middleware offers a high degree of freedom in encryption adaptiveness at a low cost.

– Previous work on selective encryption has mainly focused on how to reduce the computational cost while still making the information perceptually secure. However, the confidentiality strength in the sense of computational security has been only briefly mentioned or rudimentarily analyzed. In this thesis we address the problem of how computationally secure selectively encrypted information is by using guesswork and entropy. By adopting guesswork to selective encryption and using zero-order languages, the work reported in paper II examines when the message space is more difficult to break than the key space. In papers VI and VII, an entropy equation of selective encryption is derived and investigated for different orders of languages. To be able to generalize the entropy equation to higher dimensions, paper VIII constructs and discusses information neighborhoods. This is done by generalizing Shannon’s work on the order of languages [31] and using ideas from cellular automata [20]. The results are then applied to bitmap images.

– There is not an overwhelming amount of research that has been done for guesswork at the time of this writing, and we have therefore started to build
a theory for guesswork in this thesis. The relationship between entropy and
guesswork, using inequalities, has been examined for a time and, in the work
reported in paper III, a relationship was found between the two security mea-
ures. From the relationship, it is shown that guesswork is always greater
than or equal to entropy, with equality for the truncated geometrical distri-
bution. Moreover, as entropy is extended through joint and conditional en-
tropy, we generalize guesswork through the joint and conditional guesswork
in paper IV. In the paper it is also proven that that the joining of random
variables increases guesswork while conditioning of random variables de-
creases guesswork. Thus, adding new unknown information to the guessing
space increases guesswork and gaining information about the guessing space
decreases guesswork. This is similar to the corresponding properties of en-
tropy. In the same paper it is also proven that guesswork does not possess
the simple additive chain rule property that entropy does. Hence, no simple
way of relating the marginal, joint and conditional guesswork in a decompo-
sition equation seems possible for guesswork. Such a finding will provide a
better understanding of guesswork and the security implication of selective
encryption by making it possible to calculate guesswork from sub pieces of
the selectively encrypted information. Furthermore, time is the crucial factor
for operational security, and in paper V we generalize guesswork one step
further by investigating how guesswork changes over time through the num-
ber of incorrect guesses in multi-processor attacks. It is interesting to note
that it is possible for guesswork, as well as for entropy, to increase after an
incorrect guess. This is due to the fact that the probability distribution might
become more uniform after a guess.

9 Summary of Papers

This section summarizes the eight appended papers of the thesis. Each paper is
briefly described, addressing its position in the thesis as well as its contributions
and limitations.

Paper I – Middleware Support for Tunable Encryption

A tunable and differential treatment of security is required to achieve an appro-
priate trade-off between security and performance for wireless applications. In
this paper, we present a tunable encryption service designed as a middleware
that is based on a selective encryption paradigm. The core component of the
middleware provides block-based selective encryption. Although the selection of which data to encrypt is made by the sending application and is typically content dependent, the representation used by the core component is application and content independent. This frees the selective decryption module at the receiver of the need of application or content specific knowledge. The sending application specifies the data that shall be encrypted either directly or through a set of high-level application interfaces. A prototype implementation of the middleware is described along with an initial performance evaluation. The experimental results demonstrate that the generic middleware service offers a high degree of security adaptiveness at a low cost.

Paper II – Using Guesswork as a Measure for Confidentiality of Selectively Encrypted Messages

In this paper, we start to investigate the security implications of selective encryption. We do this by using the measure guesswork, which gives us the expected number of guesses that an attacker is assumed to perform in an optimal brute force attack to reveal the secret. The characteristics of the proposed measure are investigated for zero-order languages. We also introduce the concept of reduction chains to describe how the message (or rather search) space changes for an attacker with different levels of encryption.

Paper III – On the Relationship between Confidentiality Measures: Entropy and Guesswork

In this paper, we investigate in detail the relationship between entropy and guesswork. The aim of the study is to lay the ground for future efficiency comparison of guessing strategies. The formal definitions are given after a brief discussion of the two measures and the differences between them. A redefinition of guesswork is then given, since the measure is not completely accurate. The change is a minor modification in the last term of the sum expressing guesswork. Finally, two theorems are expressed. The first states that the redefined guesswork is equal to the concept of cross entropy, and the second states, as a consequence of the first theorem, that the redefined guesswork is equal to the sum of the entropy and the relative entropy.

Paper IV – Joint and Conditional Guesswork: Definitions and Implications

The need for computer security in today’s open computer networks is now undisputed. More and more effort is being spent on security enhancing methods and
techniques. Despite this, there is still a lack of good methods for quantitatively assessing security. New metrics that provide a more exact description of security are therefore desirable. To address this we present an in-depth investigation of the probabilistic measure guesswork, which gives the average number of guesses in an optimal brute force attack. The paper extends the definition of guesswork by defining joint and conditional guesswork. It is proved that joining increases guesswork, while conditioning reduces it. This implies that the joint guesswork is always at least equal to the marginal guesswork and that the conditional guesswork is always at most equal to the marginal guesswork. The paper also provides a description of relations and similarities between guesswork and entropy.

**Paper V – Changes in Guesswork over Time in Multi-processor Attacks**

More and more effort is being spent on security improvements in today’s computer networking environments. However, owing to the nature of computer security, there is still a lack of good quantitative assessment methods. Inventing and developing new ways of measuring security are therefore needed in order to more exactly describe, assess and improve the security of computer environments. One existing quantitative security measure is guesswork. Guesswork gives the average number of guesses in a brute force attack to succeed in breaking an encrypted message. In the current definition of guesswork, it is assumed that the attacker uses a single processor when breaking an encrypted message. An intelligent and motivated attacker will however likely use several processors that run in parallel. This paper formally investigates how guesswork changes over time in multi-processor attacks. The result is applied to three probability distributions, the English alphabet, the geometric and the truncated geometric, in order to illustrate some behaviors.

**Paper VI – Security Implications of Selective Encryption**

Quantitative measures are desirable to be able to give an analytical and more exact description of security. Two quantitative security measures that have been proposed are entropy and guesswork. When breaking an encrypted message, entropy measures the average number of guesses in an optimal binary search attack, whereas guesswork measures the average number of guesses in an optimal linear search attack. In this paper, we continue to investigate the security implications of a generic selective encryption procedure: that is, how entropy and guesswork change with the number of encrypted units, i.e., the encryption level. This is done
for languages up to the second order by deriving equations for entropy of selectively encrypted messages and then transferring the results to guesswork through an equation relating the two measures. Furthermore, unlike entropy, guesswork does not possess the chain rule, however, an equation connecting the different guessworks is derived through the equation relating entropy and guesswork.

Paper VII – Entropy of Selectively Encrypted Strings

A feature that has become desirable for low-power mobile devices with limited computing and energy resources is the ability to select a security configuration in order to create a trade-off between security and other important parameters such as performance and energy consumption. Selective encryption can be used to create this trade-off by only encrypting chosen units of the information. In this paper, we continue the investigation of the confidentiality implications of selective encryption by applying entropy to a generic selective encryption scheme. By using the concept of run-length vector from run-length encoding theory, an expression is derived for the entropy of selectively encrypted strings when the number of encrypted substrings, containing one symbol, and the order of the language changes.

Paper VIII – An Investigation of Entropy of Selectively Encrypted Bitmap Images

Selective encryption is a concept in which the main goal is to reduce computational cost while providing confidentiality by encrypting only chosen parts of the information to be protected. Previous work on selective encryption has mainly been aimed towards multimedia applications in order to reduce the overhead induced by encryption while still making the information perceptually secure to a desired protection level. This was accomplished by utilizing the fact that different parts of the information have different impacts on our perception senses, i.e., eyes and ears. How computationally secure the information is when using selective encryption has however only briefly been mentioned or rudimentarily analyzed. In this paper, we therefore investigate the security implications of selective encryption by generalizing the work on entropy of selectively encrypted strings to several dimensions and applying it to bitmap images. The generalization is done by constructing information neighborhoods that capture and model information dependencies in several dimensions.
10 Concluding Remarks and Future Work

This thesis has evaluated the performance improvement of a proposed generic selective encryption scheme. The experimental results show that the selective encryption middleware offers a high degree of freedom in encryption adaptiveness at a low cost. The confidentiality strength of selectively encrypted information in the sense of computational security when using the proposed scheme has also been investigated using entropy and guesswork. Since guesswork is less theoretically investigated than entropy and does not persist of simple additive mathematical structures, only initial results have been achieved for guesswork. However, guesswork has been adopted to selective encryption in the zero order case, and an examination has been made as to when the message space is more difficult to break than the key space of the used encryption algorithm. For entropy, an equation of selectively encrypted information has been derived for different orders and in several dimensions and applied to bitmap images. The generalization to several dimensions was done by constructing information neighborhoods from Shannon’s work on the order of languages and the neighborhood concept in cellular automata. However, more research is needed to clarify the security implications of selective encryption and, thus, be able to trade between confidentiality and performance.

To provide a better understanding of guesswork and build a theory of guesswork, this thesis extends and generalizes guesswork in several ways. The joint and conditional guesswork is defined and, as for entropy, it is proven that the joining of random variables increases guesswork while the conditioning of random variables decreases guesswork. Hence, adding information to a secret increases guesswork and gaining information about a secret decreases guesswork. This is equal to the corresponding properties of entropy. However, guesswork does not possess the chain rule decomposition property, as entropy does. Hence, decomposing guesswork into a chain rule-like equation, relating the marginal, conditional and joint guessworks, is part of future work. Such a finding will make it possible to actually calculate guesswork from sub pieces of the information, thereby making it possible to further investigate the security implications of selective encryption.

It is also proven in the thesis that guesswork is equal to the sum of the entropy and the relative entropy. Hence, guesswork is always greater than or equal to entropy, with equality for the truncated geometrical probability distribution. This finding also gives a connection between the two guessing strategies, optimal linear search and optimal binary search. Moreover, guesswork is extended to incorporate changes over time through the number of incorrect guesses in multi-
processor attacks. Normally, guesswork decreases after each incorrect guess. However, guesswork can locally increase due to the fact that the probability distribution might become more uniform after a guess. However, guesswork will decrease in the long term. Future research includes investigations of guesswork increment, i.e., the difference between two consecutive changes in guesswork values and how entropy changes in multi-processor attacks. Furthermore, in the current definition of guesswork, it is assumed that there is only one correct value in the sample space. An investigation of what happens when there is a set of correct values is also a part of future research.

References
