Integrated Encryption in Dynamic Arithmetic Compression

Shmuel T. Klein and Dana Shapira

1Bar Ilan University
2Ariel University, Israel

LATA 2017
Introduction

Concerns of Communication over a network:

1. processing speed
2. space savings of the transformed data
3. security
Introduction

Concerns of Communication over a network:

1. processing speed
2. space savings of the transformed data
3. security
Introduction

Concerns of Communication over a network:

1. processing speed
2. space savings of the transformed data
3. security
Compression Cryptosystem

- Data Compression - representation in fewer bits
- Encryption - protecting information
Compression Cryptosystem

- Data Compression - representation in fewer bits
- Encryption - protecting information
Data Compression - representation in fewer bits

Encryption - protecting information

⇒ Achieved by removing redundancies.
Compression Cryptosystem

- Data Compression - representation in fewer bits
- Encryption - protecting information

⇒ Achieved by removing redundancies.

Compression Cryptosystem
Compression Cryptosystem

- Compress then Encrypt
- Encrypt then compress
Compress then Encrypt

Encrypt then compress

Simultaneous
Why Arithmetic coding?

Huffman ???
Why Arithmetic coding?

Huffman ???

easily breakable
Why Arithmetic coding?

Huffman ???

easily breakable

communication errors
1. Arithmetic Coding
   - Static arithmetic coding
   - Adaptive Arithmetic Coding

2. Proposed Compression Cryptosystem

3. Empirical Results
   - Compression performance
   - Uniformity
   - Cryptographic attacks
Outline

1 Arithmetic Coding
   - Static arithmetic coding
   - Adaptive Arithmetic Coding

2 Proposed Compression Cryptosystem

3 Empirical Results
   - Compression performance
   - Uniformity
   - Cryptographic attacks
Arithmetic Coding

Klein & Shapira

Integrated Encryption in Dynamic Arithmetic Compression
Arithmetic Coding

Klein & Shapira

Integrated Encryption in Dynamic Arithmetic Compression
Arithmetic Coding

1.0
0.9
0.83
0.9
0.2
0.0

b

0.34
0.2

ba

b

0.326
0.3288
0.3386
0.34

Klein & Shapira
Integrated Encryption in Dynamic Arithmetic Compression
Arithmetic Coding

Static arithmetic coding
Adaptive Arithmetic Coding

Klein & Shapira
Integrated Encryption in Dynamic Arithmetic Compression
Arithmetic Coding

Adaptive Arithmetic Coding

1. compute the new interval
2. update the model by incrementing the frequency of the current character
3. adjust the relative sizes of the partition accordingly
Adaptive Arithmetic Coding

1. compute the new interval
2. update the model by incrementing the frequency of the current character
3. adjust the relative sizes of the partition accordingly
Adaptive Arithmetic Coding

1. compute the new interval
2. update the model by incrementing the frequency of the current character
3. adjust the relative sizes of the partition accordingly
Arithmetic Coding


Outline

1. Arithmetic Coding
   - Static arithmetic coding
   - Adaptive Arithmetic Coding

2. Proposed Compression Cryptosystem

3. Empirical Results
   - Compression performance
   - Uniformity
   - Cryptographic attacks
Proposed Compression Cryptosystem

Cryptosystem based on dynamic arithmetic coding

- Update the model selectively
- Use a secret key $K = k_0 k_1 \cdots k_{t-1}$
- The model is updated at step $i$ if and only if $k_{(i-1)} \mod t = 1$. 
Proposed Compression Cryptosystem

Cryptosystem based on dynamic arithmetic coding

- Update the model selectively
- Use a secret key $K = k_0 k_1 \cdots k_{t-1}$
- The model is updated at step $i$ if and only if $k_{(i-1)} \mod t = 1$. 

Klein & Shapira
Integrated Encryption in Dynamic Arithmetic Compression
Cryptosystem based on dynamic arithmetic coding

- Update the model selectively
- Use a secret key $K = k_0 k_1 \cdots k_{t-1}$
- The model is updated at step $i$ if and only if $k_{(i-1)} \mod t = 1$. 

Klein & Shapira

Integrated Encryption in Dynamic Arithmetic Compression
encode$(M, K)$
1  \( n \leftarrow |M| \)
2  \( t \leftarrow |K| \)
3  initialize the interval to be \([0, 1)\) with
4  \( \text{uniform distribution of the alphabet symbols} \)
4  for \( i \leftarrow 1 \) to \( n \)
4.1  compute the new interval
4.2  if \( k_{(i-1)} \mod t = 1 \) then
4.2.1  update the model
4.3  else
4.3.1  the new partition into intervals is the current one
5  return some value in the current interval
Compression Cryptosystem

encode($M, K$)
1 $n \leftarrow |M|$
2 $t \leftarrow |K|$
3 initialize the interval to be $[0, 1)$ with uniform distribution of the alphabet symbols
4 for $i \leftarrow 1$ to $n$
4.1 compute the new interval
4.2 if $k_{(i-1) \mod t} = 1$ then
4.2.1 update the model
4.3 else
4.3.1 the new partition into intervals is the current one
5 return some value in the current interval

$\leftarrow$ weakness

Precede plaintext by some known text
Compression Cryptosystem

encode($M, K$)
1 $n \leftarrow |M|$
2 $t \leftarrow |K|$
3 initialize the interval to be $[0, 1)$ with uniform distribution of the alphabet symbols
4 for $i \leftarrow 1$ to $n$
4.1 compute the new interval
4.2 if $k(i-1) \mod t = 1$ then
4.2.1 update the model
4.3 else
4.3.1 the new partition into intervals is the current one
5 return some value in the current interval

Precede plaintext by some known text

$\leftarrow$ weakness
Arithmetic Coding

Proposed Compression Cryptosystem

Empirical Results

Compression Cryptosystem

Precede plaintext by some known text

Klein & Shapira

Integrated Encryption in Dynamic Arithmetic Compression

\begin{verbatim}
encode(M, K)
1  n ← |M|
2  t ← |K|
3  initialize the interval to be [0, 1) with uniform distribution of the alphabet symbols
4  for i ← 1 to n
4.1  compute the new interval
4.2  if k_{(i-1) mod t} = 1 then
4.2.1  update the model
4.3  else
4.3.1  the new partition into intervals is the current one
5  return some value in the current interval
\end{verbatim}
Worst Case Example: abababab...
Worst Case Example: abababab· · ·

Dynamic AC: uniform probability distribution \((\frac{1}{2}, \frac{1}{2})\) for any even sized history window

Selected: The probability of uniform distribution:

\[
\frac{\binom{n}{n/2} \binom{n}{n/2}}{\binom{2n}{n}} \sim \frac{2}{\sqrt{\pi n}},
\]

by Stirling’s approximation.
Compression Efficiency

Worst Case Example: abababab\ldots

Dynamic AC: uniform probability distribution \((\frac{1}{2}, \frac{1}{2})\) for any even sized history window

Selected: The probability of uniform distribution:

\[
\frac{\left(\frac{n}{2}\right)^2}{\binom{2n}{n}} \sim \frac{2}{\sqrt{\pi n}},
\]

by Stirling’s approximation.

\(\rightarrow\) For a key of size 512, only in 5\% of the cases will the exactly uniform model be obtained.
Outline

1. Arithmetic Coding
   - Static arithmetic coding
   - Adaptive Arithmetic Coding

2. Proposed Compression Cryptosystem

3. Empirical Results
   - Compression performance
   - Uniformity
   - Cryptographic attacks
Empirical Results

Data Sets

1. \textit{ebib} - the Bible (King James version) in English
2. \textit{ftxt} - the French version of the European Union’s JOC corpus, a collection of pairs of questions and answers on various topics used in the ARCADE evaluation project
3. \textit{sources} - formed by C/Java source codes obtained by concatenating .c, .h and .java files of the linux-2.6.11.6 distributions
4. \textit{English} - the concatenation of English text files selected from the collections of the Gutenberg Project
**Empirical Results**

<table>
<thead>
<tr>
<th>File</th>
<th>full size (MB)</th>
<th>compressed size (MB)</th>
<th>absolute loss (bytes)</th>
<th>relative loss ($\times 10^{-5}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ftxt</td>
<td>7.6</td>
<td>4.2</td>
<td>316</td>
<td>7</td>
</tr>
<tr>
<td>sources</td>
<td>200.0</td>
<td>136.6</td>
<td>436</td>
<td>3</td>
</tr>
<tr>
<td>English</td>
<td>1024.0</td>
<td>579.3</td>
<td>437</td>
<td>7</td>
</tr>
</tbody>
</table>
## Empirical Results

### Compression performance

<table>
<thead>
<tr>
<th>File</th>
<th>Full size (MB)</th>
<th>Compressed size (MB)</th>
<th>Absolute loss (bytes)</th>
<th>Relative loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>ftxt</td>
<td>7.6</td>
<td>4.2</td>
<td>316</td>
<td>$7 \times 10^{-5}$</td>
</tr>
</tbody>
</table>

Klein & Shapira: Integrated Encryption in Dynamic Arithmetic Compression
## Empirical Results

### Compression performance

<table>
<thead>
<tr>
<th>File</th>
<th>full size</th>
<th>compressed size</th>
<th>absolute loss</th>
<th>relative loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>ftxt</td>
<td>7.6</td>
<td>4.2</td>
<td>316</td>
<td>$7 \times 10^{-5}$</td>
</tr>
<tr>
<td>sources</td>
<td>200.0</td>
<td>136.6</td>
<td>436</td>
<td>$3 \times 10^{-6}$</td>
</tr>
</tbody>
</table>
## Empirical Results

### Compression performance

<table>
<thead>
<tr>
<th>File</th>
<th>Full size (MB)</th>
<th>Compressed size (MB)</th>
<th>Absolute loss (bytes)</th>
<th>Relative loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>ftxt</td>
<td>7.6</td>
<td>4.2</td>
<td>316</td>
<td>$7 \times 10^{-5}$</td>
</tr>
<tr>
<td>sources</td>
<td>200.0</td>
<td>136.6</td>
<td>436</td>
<td>$3 \times 10^{-6}$</td>
</tr>
<tr>
<td>English</td>
<td>1024.0</td>
<td>579.3</td>
<td>437</td>
<td>$7 \times 10^{-7}$</td>
</tr>
</tbody>
</table>
Uniformity

Probability of occurrence of substrings as function of value
## Uniformity

Probability of occurrence of substrings as function of value

<table>
<thead>
<tr>
<th>value</th>
<th>standard arithmetic</th>
<th>selective updates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$m = 3$</td>
<td>$m = 3$</td>
</tr>
<tr>
<td>0</td>
<td>0.12503</td>
<td>0.12507</td>
</tr>
<tr>
<td>1</td>
<td>0.12498</td>
<td>0.12503</td>
</tr>
<tr>
<td>2</td>
<td>0.12510</td>
<td>0.12491</td>
</tr>
<tr>
<td>3</td>
<td>0.12499</td>
<td>0.12499</td>
</tr>
<tr>
<td>4</td>
<td>0.12498</td>
<td>0.12503</td>
</tr>
<tr>
<td>5</td>
<td>0.12511</td>
<td>0.12488</td>
</tr>
<tr>
<td>6</td>
<td>0.12499</td>
<td>0.12499</td>
</tr>
<tr>
<td>7</td>
<td>0.12482</td>
<td>0.12499</td>
</tr>
<tr>
<td></td>
<td>$m = 2$</td>
<td>$m = 2$</td>
</tr>
<tr>
<td>0</td>
<td>0.25002</td>
<td>0.25010</td>
</tr>
<tr>
<td>1</td>
<td>0.25009</td>
<td>0.24991</td>
</tr>
<tr>
<td>2</td>
<td>0.25009</td>
<td>0.24999</td>
</tr>
<tr>
<td>3</td>
<td>0.24981</td>
<td>0.25009</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.12503</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.12488</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0.12499</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>0.12499</td>
</tr>
<tr>
<td></td>
<td>$m = 1$</td>
<td>$m = 1$</td>
</tr>
<tr>
<td>0</td>
<td>0.50011</td>
<td>0.500005004</td>
</tr>
<tr>
<td>1</td>
<td>0.49989</td>
<td>0.499994996</td>
</tr>
<tr>
<td>2</td>
<td>0.12491</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.12499</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.12503</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.12488</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.12499</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.12499</td>
<td></td>
</tr>
</tbody>
</table>
Uniformity

Ratio $\frac{\sigma}{\mu}$ of standard deviation to average within the set of $2^m$ values for $m = 1, \ldots, 8$.

<table>
<thead>
<tr>
<th></th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard</td>
<td>0.00383</td>
<td>0.00251</td>
<td>0.00164</td>
<td>0.00125</td>
<td>0.00094</td>
<td>0.00072</td>
<td>0.00053</td>
<td>0.00030</td>
</tr>
<tr>
<td>selective</td>
<td>0.00135</td>
<td>0.00042</td>
<td>0.00207</td>
<td>0.00182</td>
<td>0.00059</td>
<td>0.00013</td>
<td>0.0003</td>
<td>0.00001</td>
</tr>
</tbody>
</table>
Cryptographic attacks

Overlapping intervals

```
0  a  b  c  d  e  1
0  a  b  c  d  e  1
```

Cumulative size of boldfaced sub-intervals = 0.714

Klein & Shapira
Integrated Encryption in Dynamic Arithmetic Compression
Cryptographic attacks

Overlapping intervals

Cumulative size of boldfaced sub-intervals = 0.714
Cryptographic attacks

Size of overlapping intervals as a function of the number of processed characters.
Cryptographic attacks

Size of overlapping intervals as a function of the number of processed characters.

Probability for a correct guess after 10 characters $\leq 0.00006$
sensitivity to variations in the secret key

The Normalized Hamming distance: Let $A = a_1 \cdots a_n$ and $B = b_1 \cdots b_m$ be two bitstrings and assume $n \geq m$.

The normalized Hamming distance: $\frac{1}{n} \sum_{i=1}^{n} (a_i \text{ XOR } b_i)$. 
sensitivity to variations in the secret key

Normalized Hamming distance

Klein & Shapira
Integrated Encryption in Dynamic Arithmetic Compression
Thank You!