Forensic Behavior Analysis in Video Conferencing based on the Metadata of encrypted Audio and Video Streams - Considerations and Possibilities

Robert Altschaffel, Jonas Hielscher, Kevin Lamshöft, Christian Krätzer, Jana Dittmann

Email: Robert.Altschaffel@iti.cs.uni-magdeburg.de

Otto-von-Guericke University
Magdeburg, Germany
• Research Assistant in Research Group Multimedia and Security, Otto-von-Guericke-University of Magdeburg
• Research interests:
  – Computer Forensics
  – Automotive IT
  – ICS (Industrial Control Systems)
  – Network Analysis
  – Data Protection/Privacy
• Broad range of publications on these research subjects
• Research group at the Otto-von-Guericke University Magdeburg, Germany
• Research fields and interests
  – Computer security, privacy, data sovereignty
  – Security in Automotive IT and Industrial Control systems (ICS)
  – Forensics (Desktop IT, crime scene, Automotive IT, Industrial Control Systems)
  – Watermarking and Steganography
  – Biometrics
• https://omen.cs.uni-magdeburg.de/itiamsl/deutsch/home/index.html
Outline

• Introduction
• State of the art
  – Activity and content identification in encrypted traffic
  – Computer forensics
• Usability of audio and video streams for activity analysis
• Exemplary implementation and preliminary results
  – Test setup
  – Pattern Recognition
  – Results
• Conclusion
  – Future Work
• **Video Conferencing (VC)** is of increased importance during these times of crisis
• Video Conferencing often includes
  – Video Channel
  – Audio Channel
  – Text Channel
• Communication is usually encrypted – but **what can be observed despite the encryption?**
• Paper focuses on VC relying on a **central communication server**
State of the art: Activity and content identification in encrypted traffic

- Various work on identifying activities or content in encrypted network traffic
  - Activities which are transmitted ‘live’
  - Basic Idea: Different activities led to different communication behavior
- Examples:
  - Reconstruction of inputs during a SSH session based on packet sizes and inter-packet times [1]
  - Identification of activities during a TeamViewer session based on properties of the network traffic [2]
  - Reconstruction of conversation in encrypted Skype traffic based on the size of transmitted packets and timing information [3]
  - Identification of speakers in Skype session based packet size and timing [4]

State of the Art: Computer Forensics

- Forensics describes a scientific and systematic approach for the reconstruction of events
- Forensic Process Models support the forensic process
  - Structuring the process
  - Making the process easier to describe and compare
- In this paper we use the Forensic Process Model from [1]
  - Of benefit for this paper:
  - Structures the forensic process into
    - 6 Investigation Steps (phases of the process including a Strategic Preparation)
    - 8 Data Types (describing how certain data is handled during the forensic process)

➤ Aim: identify a **structured and comparable** approach for activity identification during VC

Usability of audio and video streams for activity analysis

• General approach:
  – Identify the various activities which might influence communication behavior
  – Identify which properties of the communication behavior are influenced by different activities
  – Identify where these properties can be observed
  – Create an overall process
Usability of audio and video streams for activity analysis: Activities

• Identifying activities which lead to differences in communication

  **Activities in Text**
  – $TE_1$ inactive / not typing
  – $TE_2$ typing
  – $TE_3$ sending text

  **Activities in Audio**
  – $A_1$ deactivated / muted
  – $A_2$ unmute and silent
  – $A_3$ unmute and speaking fluently
  – $A_4$ unmute and speaking chopped off

  **Activities in Video**
  – $V_1$ deactivated
  – $V_2$ black screen
  – $V_3$ one person in front
  – $V_4$ multiple persons in front
Usability of audio and video streams for activity analysis: Properties

• Properties based on packet size and timing (a used in [1], [2], [3] and [4])
• Features are extracted from these properties by an feature extractor based on the work in [2]
  ➢ Window-based features using fixed time windows
    – Packet size (minimum, maximum, mean, deviation)

Usability of audio and video streams for activity analysis: Availability within infrastructure

- Different systems take part in enabling VC
  - Various clients (CL\(_1\)–\(_3\))
  - A central server (CS)
  - Network Infrastructure (NI)

- Observable Communication differs at various points
  - Also in terms of accessible data (Data Types from [1] in the extension from [2])
  - DT2 = raw, not interpreted data
  - DT3 = meta data
  - DT9 = interpreted audio/video stream

Usability of audio and video streams for activity analysis: Process

- **Pattern Recognition** is used to identify various activities which can be mapped to the **Investigation Steps** from [1]
  - Training of a decision model before the classification takes place (= Strategic Preparation)
  - Model can then be used to classify gathered data (= Data Analysis/Data Investigation)
  - This data has to be gathered before (= Data Gathering)
  - The entire process is documented (= Documentation)

---

Exemplary implementation and preliminary results

• Tests with Zoom [1] and BBB [2]
• Data Acquisition by test setup
• Pre-Processing and Feature Extraction based on [3]
• Model Generation and Classification done using WEKA [4]
• Visual confirmation of Pattern Recognition results

Exemplary implementation and preliminary results

- Test of three different solutions
  - BBB [1]
  - Zoom-App [2]
  - Zoom-Web [2]
- Data Acquisition at $O_6$
  - Capturing only incoming traffic at a passive observer
- Test of activities
  - In Text
  - In Audio
  - In Video
- Goal: Distinguish user behavior
  - Visual verification
  - Classifier (Pattern Recognition)

Only one extraction point at a passive observer is used in our test setup.

Exemplary implementation and preliminary results – Test cases

- **[T1]** - Audio
  - CL1 is using the microphone to send audio
  - Different levels of audio usage are compared
    - A1 microphone is muted in the conference client and on the hardware
    - A2 microphone is activated in the conference client but deactivated on the hardware
    - A3 microphone is fully activated and a monotone voice is recorded
    - A4 microphone is fully activated and a voice which varies in vocal pitch and volume is recorded
- **[T2]** - Video:
  - CL1 is using the built-in webcam to send video data
  - Different levels of video usage are compared
    - V1 The webcam is deactivated in the client
    - V2 The webcam is activated and a black image is recorded
    - V3 The webcam is activated and a static video (without visible movement) is recorded
    - V4 The webcam is activated and a moving video (movement of a person) is recorded
- **[T3]** - Video2x
  - CL1 and CL2 are using the built-in camera and both perform tests like in [T2] (V5)
  - test whether an observer can identify the number of active participants or not
- **[T4]** - Video-Audio
  - CL1 is using different audio- and video features like described in [T1] and [T2]
  - aim is to test whether the stream of audio and video data can be separated on network level in order to evaluate them separately
Exemplary implementation and preliminary results – Pattern Recognition

- Training of classifier with WEKA [1] with J48 algorithm
  - Success in 9 out of twelve performed tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Zoom-Web</th>
<th>Zoom-App</th>
<th>BBB</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>0.9989</td>
<td>0.9947</td>
<td>1</td>
</tr>
<tr>
<td>T2</td>
<td>0.9993</td>
<td>0.9953</td>
<td>1</td>
</tr>
<tr>
<td>T3</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>T4</td>
<td>0.9993</td>
<td>0.9947</td>
<td>NA</td>
</tr>
</tbody>
</table>

Kappa statistics (in the range [0,1] with a value of 1 indicating optimal classification) for the different test cases.

<table>
<thead>
<tr>
<th>Classified as</th>
<th>( V_1 ) deactivated</th>
<th>( V_2 ) black screen</th>
<th>( V_3 ) one person in front</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V_1 ) deactivated</td>
<td>44</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( V_2 ) black screen</td>
<td>1</td>
<td>1221</td>
<td>0</td>
</tr>
<tr>
<td>( V_3 ) one person in front</td>
<td>1</td>
<td>0</td>
<td>2898</td>
</tr>
</tbody>
</table>


Exemplary implementation and preliminary results – Visual Verification

- Visual verification
  - Succeeded in most cases

In case of the zoom app, different audio usage can be clearly distinguished by the amount of incoming (UDP) traffic at the passive observation point.
Exemplary implementation and preliminary results

- Visual verification
  - Succeeded in most cases
- [T1]: Audio
- [T2]: Video
- [T3]: Video2x
- [T4]: Video-Audio

<table>
<thead>
<tr>
<th>Test</th>
<th>Zoom-Web</th>
<th>Zoom-App</th>
<th>BBB</th>
</tr>
</thead>
<tbody>
<tr>
<td>[T1]</td>
<td>$A_1 / A_2 / A_3 / A_4$</td>
<td>$A_1 / A_2 / A_3 / A_4$</td>
<td>$(A_1 \land A_2) / (A_3 \land A_4)$</td>
</tr>
<tr>
<td>[T2]</td>
<td>$V_1 / V_2 / V_3$</td>
<td>$V_1 / V_2 / V_3$</td>
<td>$V_1 / (V_2 \land V_3)$</td>
</tr>
<tr>
<td>[T3]</td>
<td>$V_1 / V_2 / V_3 / A_1 / A_2 / A_3 / A_4$</td>
<td>$V_1 / V_2 / V_3 / A_1 / A_2 / A_3 / A_4$</td>
<td>$X$</td>
</tr>
<tr>
<td>[T4]</td>
<td>$X$</td>
<td>$V_1 / V_2 / V_3 / V_4$</td>
<td>$(V_1 \land V_2 \land V_3) / V_4$</td>
</tr>
</tbody>
</table>

In the twelve tests, different usage of audio and video data can be distinguished from each other by simple visual verification of the I/O graph.
Summary

- Identification of various activities within encrypted audio/video streams during Video Conferencing seems feasible
- Systematic approach based on Pattern Recognition and forensic principles
- Clear definition of the various possible points to observe VC communication and their impact on forensic investigations

- Future aspects
  - Extend training data set (in terms of used systems, number of users, observation points, etc.)
  - Potential use of biometrics to identify persons within these encrypted audio/video streams

The research shown in this paper is partly funded by the European Union Project "CyberSec LSAOVGU-AMSL"