TIE
Traffic Identification Engine

Alberto Dainotti - alberto@unina.it
COMICS Research Group
University of Napoli “Federico II”
TRAFFIC CLASSIFICATION

To associate *flows* to the *applications* that generate them

\{ \text{UDP, IP}_{\text{SRC}}: 10.0.0.1, \text{PORT}_{\text{SRC}}: 31215, \text{IP}_{\text{DST}}: 212.48.72.19, \text{PORT}_{\text{DST}}: 80} \}

\Rightarrow \text{SKYPE!}

\{ \text{TCP, IP}_{\text{SRC}}: 10.0.0.1, \text{PORT}_{\text{SRC}}: 2233, \text{IP}_{\text{DST}}: 13.29.10.199, \text{PORT}_{\text{DST}}: 25} \}

\Rightarrow \text{SMTP!}
MOTIVATIONS

Why classify traffic?

• **To understand** what our links carry
  - How are people using the Internet?
  - What’s the killer application?
  - Does it really matter to model this or that?
  - Is something “strange” happening and we don’t know it?

• **To operate** networks
  - Resource allocation and QoS
  - Enforcement of security policies (e.g. Firewalling)
  - Billing based on typology of traffic
  - Network provisioning
  - Diagnostics: retracing phenomena (e.g. congestion) to specific applications and protocols
APPROACHES

an evolving complex scenario

• **Port-based**
  ✓ Fast and Simple
  - Unreliable (e.g. TCP:80 ≠ HTTP)

• **Payload inspection**
  ✓ Often reliable
  - Privacy concerns
  - Computationally heavy
  - Can be tricked by protocol encapsulation, encryption, ...

• **Pattern Recognition & Behavioral**
  ✓ Promising with respect to current trends (encryption, obfuscation, novel applications, …)
  - Experimental
  - Reliable?

SCIENCE EFFORTS

dramatically increased in past years

- Tools for pattern matching on packets payload become common (in the IDS field)
- RFC 1340 – Reynolds, Postel – establishes the registered port space
- Peer-to-Peer traffic hiding behind well-known ports is identified through payload inspection
- Machine Learning techniques are proposed for traffic classification

Cumulative sum of papers on IEEEExplore matching “traffic classification” OR “traffic identification” (filtered on “Communication, Networking”)
WHERE WE ARE

difficulties...

• A lot of work is still in experimental stage
• Scarce availability of real implementations
• Sharing traffic data in scientific community
• Lack of benchmarks
• Lack of standard formats
WHERE WE ARE

... and opportunities

• Large interest of different communities
  - Scientists
  - Providers
  - Industry
  - Society

• Several approaches and code proved to be effective

• Increasing complexity of Internet applications and traffic will continue to keep this topic hot!
TIE
Traffic Identification Engine

A software **platform** for **building** traffic classifiers and for **experimenting** with them

- Multi-approach Framework
- Open source
- Fast (*C* language, *Libpcap*, *Endace DAG* support, ...)
- Modular
- Supports **multiclassification**
- Supports **online** traffic classification
- ....
TIE HISTORY

the genesis

Started in 2007 by researchers of the “TRAFFIC” project inside COMICS

http://tie.comics.unina.it

COMICS Research Group
University of Napoli “Federico II” - Italy
TIE HISTORY
opening to the world

During these 4 years has been/is the subject of
• Graduate and undergraduate students theses
• Collaborations with other research groups

• Collaborations with the Industry (manufacturing, customer service assurance consultancy, ...)
• National and European Research Projects

COMICS Research Group
University of Napoli “Federico II” - Italy
Papers

- V. Carela-Espanol, P. Barlet-Ros, M. Solé-Simò, A. Dainotti, W. de Donato, A. Pescapè, K-dimensional trees for continuous traffic classification, International Workshop on Traffic Monitoring and Analysis (TMA’10) @ PAM 2010 - April 2010, Zurich (Switzerland)

Book Chapters


Technical Reports


Patents

• **Offline**
  - a session is *classified* only when it ends or at the end of TIE execution

• **Realtime**
  - a session is classified as soon as possible and output is immediately available

• **Cyclic**
  - the classification of all live sessions is generated at regular intervals (e.g. each 5 min.)

---

In Cyclic mode, automated Web Reports can be built using CAIDA’s CoralReef tools.
• It can work with configurable definitions of sessions
  - **Flows**
    - \( \langle L4Proto, IP_{src}, Port_{src}, IP_{dst}, Port_{dst} \rangle + \text{timeout} \)
  - **Biflows**
    - Same as above but src and dst are swappable
    - Support for TCP connections through simple heuristics based on TCP flags
  - **Hosts**
    - Under development
# CLASSIFICATION PLUGINS

<table>
<thead>
<tr>
<th>Name</th>
<th>Based on</th>
<th>Status</th>
<th>Contributor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Port</td>
<td>L4 ports</td>
<td>Available</td>
<td>UNINA (signatures from CAIDA)</td>
</tr>
<tr>
<td><strong>L7</strong></td>
<td>Deep Payload Inspection</td>
<td>Available</td>
<td>UNINA (signatures/code from Linux L7-filter)</td>
</tr>
<tr>
<td>PortLoad</td>
<td>Lightweight Payload Insp.</td>
<td>Licensable</td>
<td>UNINA</td>
</tr>
<tr>
<td>GMM-PS</td>
<td>Statistical Approach: PS</td>
<td>Under Test</td>
<td>UNINA</td>
</tr>
<tr>
<td>HMM</td>
<td>Statistical Approach: PS, IPT</td>
<td>Under Test</td>
<td>UNINA</td>
</tr>
<tr>
<td>FPT</td>
<td>Statistical Approach: PS, IPT</td>
<td>Under Dev</td>
<td>UNIBS</td>
</tr>
<tr>
<td>Joint</td>
<td>Machine Learning: PS, IPT</td>
<td>Under Test</td>
<td>UNINA-CENS</td>
</tr>
<tr>
<td>GT</td>
<td>Information from hosts</td>
<td>Under Dev</td>
<td>UNINA-UNIBS</td>
</tr>
<tr>
<td>OpenDPI</td>
<td>Deep Payload Inspection</td>
<td>Beta</td>
<td>OpenDPI, UNINA, TUM</td>
</tr>
<tr>
<td><strong>WEKA</strong></td>
<td>Imports the output of a WEKA classifier</td>
<td>Available</td>
<td>UNINA</td>
</tr>
</tbody>
</table>
• A set of utilities is distributed with TIE for the post-processing of the output
• In *realtime* mode, the output can also be sent through network sockets to another application
A CASE STUDY

PortLoad*

• TIE’s modular framework allows to easily implement a new classification technique and run it on real traffic.

• By using a unified framework and standard definitions and formats it is easy to compare and benchmark three different classification techniques.


Patent pending “Method and system for traffic classification in communication networks using content-based signatures”. 9th March 2010 - code NA2010A000011#
PORTLOAD
merging two “worlds” in traffic classification

Port-based approach
- Very inaccurate
+ Simple & Fast
+ Privacy-friendly

Deep Packet Inspection
+ Accurate
- CPU intensive
- Doesn’t care about Privacy
PORTLOAD

do we need all that payload?

• Experiments on sample traces with TIE-L7 (L7-Filter DPI based on regular expressions)
  - Evaluated where the matches happen
  - Packet position inside flow
  - Bytes in payload

• E.g.
  - 87% of the matches start at the first packet
  - Almost all matching strings start (99.98%) and finish (90.77%) in the first 32 bytes of payload of a packet

<table>
<thead>
<tr>
<th>Site</th>
<th>Date</th>
<th>Size</th>
<th>Pkts</th>
<th>biflows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univ. Napoli</td>
<td>Oct 3rd 2009</td>
<td>59 GB</td>
<td>80M</td>
<td>1M</td>
</tr>
</tbody>
</table>
PORTLOAD

taking the benefits of both approaches

• **Port-based** is *fast* and *privacy-friendly* because:
  - It needs the 1st packet only
  - It uses fixed fields (protocol and port)
  - It uses few data
  - It can be considered as a special case of packet-classification techniques developed for routers, flow-monitors, etc.

• **Payload-based** is *accurate* because it relies on application-level headers and other information from the payload
  - Payload-based signatures
PORTLOAD

$Port + Payload = PortLoad$

- A single packet (1st one with payload), fixed fields, limited data (e.g. 32B of payload)

- Payload-based signatures
  - Example of signature for the Shoutcast MP3 streaming application
    
    | App.ID | TCP/UDP | direction | offset | fields |
    |--------|---------|-----------|--------|--------|
    | 34     | UDP     | BOTH      | 0      | I      |

- Packet-classification matching approach
  - Independent field searches
  - E.g. bitmap intersection (Lakshman and Stiliadis, SIGCOMM Computer Communication Review, 1998)
A **bitmap** is assigned to each Field-Value pair

1’s in a bitmap indicate signatures compatible with that pair

AND-ing the bitmaps corresponding to packet content will return the matching signatures
PORTLOAD
evaluation of classification accuracy

- Evaluation (accuracy against TIE-L7) on UNINA trace from Oct. 2009, with a preliminary set of signatures
- We compared results on the same traffic trace obtained with
  - TIE-L7
  - TIE-PortLoad
  - TIE-Port

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy on applications sessions</th>
<th>Accuracy on applications bytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>PortLoad</td>
<td>74.24%</td>
<td>97.83%</td>
</tr>
<tr>
<td>Port-based</td>
<td>19.57%</td>
<td>25.12%</td>
</tr>
</tbody>
</table>
PORTLOAD

evaluation of performance

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Mean Time (µsec)</th>
<th>Mean Time (vs Port-based)</th>
<th>Variance (µsec^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Port-based</td>
<td>2.48</td>
<td>1.0</td>
<td>0.88</td>
</tr>
<tr>
<td>PortLoad</td>
<td>6.99</td>
<td>2.8</td>
<td>11.15</td>
</tr>
<tr>
<td>L7-Filter</td>
<td>211.4</td>
<td>85.2</td>
<td>47057.88</td>
</tr>
</tbody>
</table>
TIE DEPLOYMENT

what do you need at least

• A **Linux/FreeBSD** box
• An **optical splitter** or switch/router doing **port mirroring**
• A spare **network adapter** or an **ENDACE DAG card**
• The **pcap** library
• The CAIDA’s **CoralReef** library for live web reports

E.g. we live monitor a 200Mbps link with a Xeon box / FreeBSD 6.3 and a ~$800 DAG card.
RIPE MEETING
TIE and Internet Service Providers

• We are always seeking for **collaborations**

TIE can be used by ISPs for:
- Deploying traffic classification with **low costs**
- Developing traffic classifiers targeted to specific needs and **operating problems** (novel/custom network protocols and encapsulations, specific classes of traffic and applications, etc.)
- Helping in **monitoring and diagnosing** network events
- Deploy differentiating **QoS** or **security** policies
- **Forecasting** users-traffic trends
- ...

• We are particularly interested in **listening to ISPs needs** and unsolved technical problems and to discuss them
THANKS

feedback is very welcome

http://www.grid.unina.it/Traffic

This work has been funded by the European Project “OneLab2” (ICT FP7 IP 224263).