Degree-based Outlier Detection within IP Traffic Modelled as a Link Stream

Audrey Wilmet\(^1\), Tiphaine Viard\(^1\), Matthieu Latapy\(^1\), Robin Lamarche-Perrin\(^2\)

\(^1\)Laboratoire d’informatique de Paris 6 (LIP6) - Complex Networks Team [http://www.complexnetworks.fr/](http://www.complexnetworks.fr/)
\(^2\)Institut des Systèmes Complexes Paris Île de France (ISC-PIF)

TMA Conference 2018, Vienna, Austria
June 26-29, 2018
**Context and Goals**

Detect outliers, identify their cause, remove them from IP traffic:

![Graph showing number of distinct IP addresses over time](image-url)
Context and Goals

Detect outliers, identify their cause, remove them from IP traffic:

![Graph showing number of distinct IP addresses over time](image)
IP Traffic as a Link Stream

Link stream constructed from 1h of IP Traffic (MAWI):
- **Nodes** = IP addresses
- **Interactions** = packet exchanges
- **Link stream construction:**
  
  *Two nodes are linked together from time $t_1$ to time $t_2$ if they exchanged at least one packet every second within this time interval.*

![Diagram showing nodes a, b, and c interacting over time](image)

- **ex:** nodes a and c interact from $t_1 = 3$ to $t_2 = 4$

Degree of \((v, t)\)

\[ d_t(v) = \text{Number of neighbours of node } v \text{ at time } t \]

**Example: degree profile of b**

![Diagram showing degree profiles of nodes a, b, and c over time]

Audrey Wilmet, Tiphaine Viard, Matthieu Latapy, Robin Lamarche-Perrin
Our Approach
Detection: example

- Detection:
  Find observations of the degree which deviate statistically from others.

![Diagram showing detection example]
Detection: example

- **Detection:**
  Find observations of the degree which deviate statistically from others.

Degree distribution on all couples \((v, t)\):
Detection: example

Find observations of the degree which deviate statistically from others.

Degree distribution on all couples \((v, t)\):

Detected outlier:

\[ d_t(v) = 7 \]
Identification: example

- Identification:
  Find entities which are responsible for the outlying degree observation.

Detected outlier:
⇒ outlying observation $d_t(v) = 7$.

Identified outlier:
Identification: example

- Identification:
  Find entities which are responsible for the outlying degree observation.

Detected outlier:
⇒ outlying observation $d_t(v) = 7$.

Identified outlier:
⇒ the set : $\{(e, t) \mid t \in [20, 21]\}$
Removal: example

Remove identified entities from the link stream.

Detected outlier:
⇒ outlying observation $d_t(v) = 7$.

Identified outlier:
⇒ the set : $\{(e, t) \mid t \in [20, 21]\}$

Removed outlier:
3 Removal: example

- **Removal:**
  Remove identified entities from the link stream.

**Detected outlier:**
⇒ *outlying observation* $d_t(v) = 7$.

**Identified outlier:**
⇒ *the set*: $\{(e, t) \mid t \in [20, 21]\}$

**Removed outlier:**
⇒ $\{(e, t) \mid t \in [20, 21]\}$
Detection in our data

Link stream constructed from 1h of IP Traffic (MAWI)

Degree distribution on all couples \((v, t)\):
Difficulties

Outlier = Activity that deviates from the usual one

Find an outlier ⇔ Find the normality

Heterogeneous

Homogeneous with outliers
Our Method
Local Degree Distributions

Degree observation on substreams with a duration of 2 seconds.
Local Distributions Similarity

Audrey Wilmet, Tiphaine Viard, Matthieu Latapy, Robin Lamarche-Perrin
Comparison of Local Distributions

**Introduction**

**Context and Goals**

**Link Stream**

**Degree**

**Our Approach**

In theory

In Practice

Difficulties

**Our Method**

Distributions

Similarity

Detection

Identification

Removal

Validation

**Conclusion**

Audrey Wilmet, Tiphaine Viard, Matthieu Latapy, Robin Lamarche-Perrin
Comparison of Local Distributions

Degree Class $C_j$ Distribution

$P(C_{10}) = \{16, ..., 20\}$

$P(C_1) = \{1\}$

Distribution on $T_i$
Results: Homogeneous Distributions

Audrey Wilmet, Tiphaine Viard, Matthieu Latapy, Robin Lamarche-Perrin
Results: Homogeneous Distributions

There are a lot more couples $(v, t)$ for which $d_t(v) \in C_2$ during $T_9$ than during the majority of other time slices.
Overview of our method

Comparison of local distributions

Detection of outliers
Identification: difficulties

Detected Outlier $= 2$ informations
$\Rightarrow$ time slice $T_i$ + degree class $C_j$

How to find responsible entities?
How to identify detected outliers?

Previous example:
Detected Outlier $\Rightarrow T_9$ and $C_2$

Difficulty: there are numerous $(v, t)$ within $C_2$ during $T_9$

Which of them are abnormal?
Identification

- **Low degree classes:**
  
  \[ \text{outlier} = \text{normal} + \text{abnormal traffic} \]

- **High degree classes:**
  
  \[ \text{outlier} = \text{abnormal traffic only} \]

→ **Direct identification possible in high degree classes only**
Identification in High Degree Classes

\[ C_{41} \]

Distribution on \( T_i \)

Number of \((v, t)\) s.t. \( d_t(v) \in C_{41} \)

Detected Outlier:
\[ T_{335} = [710, 712] \text{ and } C_{41} \]

Identified outlier:
\[ \{(v_1, t) | t \in [710.3, 713.1] \} \]

\( v_1 \) is the only node having a degree within \( C_{41} \) during \( T_{335} \)
Overview of our method

Comparison of local distributions

Detection of outliers

Identification in high degree classes
Removals of identified outliers

Disappearance of the detected outlier in the $C_{41}$ distribution:

$$\{(v_1, t) \mid t \in [710.3, 713.1]\}$$
Removals of identified outliers

Disappearance of the detected outlier in the $C_{41}$ distribution:

... as well as in a smaller degree class distribution:

⇒ Allows to identify low degree classes outliers among neighbours of the removed node.
Overview of our method

Comparison of local distributions

Detection of outliers

Identification in high degree classes

Iterative removal

Identification in low degree classes

Removal

Audrey Wilmet, Tiphaine Viard, Matthieu Latapy, Robin Lamarche-Perrin
Overview of our method

Comparison of local distributions

Detection of outliers

Identification in high degree classes

Identify outliers

Iterative removal

Identification in low degree classes

Removal

Validation

Audrey Wilmet, Tiphaine Viard, Matthieu Latapy, Robin Lamarche-Perrin
Degree Classes Distributions After Removals

Disappearance of most outliers without creation of negative outliers.
Degree Classes Distributions After Removals

Disappearance of most outliers without creation of negative outliers.
Creation of normal traffic

Number of detected outliers: 1,358
Number of identified outliers: 1,163 = 85% of the detected outliers

⇒ Consequence on the number of distinct IP addresses per second.
Creation of normal traffic

Number of detected outliers: 1,358
Number of identified outliers: 1,163 = 85% of the detected outliers

⇒ Consequence on the number of distinct IP addresses per second.
Creation of normal traffic

Number of detected outliers: 1,358
Number of identified outliers: 1,163 = 85% of the detected outliers

⇒ Consequence on the number of distinct IP addresses per second.
Creation of normal traffic

Number of detected outliers: 1,358
Number of identified outliers: 1,163 = 85% of the detected outliers

⇒ Consequence on the number of distinct IP addresses per second.
Creation of normal traffic

Number of detected outliers: 1,358
Number of identified outliers: 1,163 = 85% of the detected outliers

⇒ Consequence on the number of distinct IP addresses per second.
Creation of normal traffic

Number of detected outliers: 1,358  
Number of identified outliers: 1,163 = 85% of the detected outliers

⇒ Consequence on the number of distinct IP addresses per second.
Conclusion
Conclusion

Design of a method to detect and precisely identify outliers in heterogeneous distributions:

– Structural and temporal similarity evaluation of distributions.
– Modelling of IP traffic as a link stream.

→ IP with anomalous degree profile, network scans.
Conclusion

Design of a method to detect and precisely identify outliers in heterogeneous distributions:

– Structural and temporal similarity evaluation of distributions.
– Modelling of IP traffic as a link stream.

→ IP with anomalous degree profile, network scans.

Iterative removal of identified outliers

→ Validation: Creation of normal traffic (w.r.t $d_t(v)$).
Conclusion

Design of a method to detect and precisely identify outliers in heterogeneous distributions:
- Structural and temporal similarity evaluation of distributions.
- Modelling of IP traffic as a link stream.
  - IP with anomalous degree profile, network scans.

Iterative removal of identified outliers
  - Validation: Creation of normal traffic (w.r.t $d_t(v)$).

⇒ Method applicable over temporal interactions in general.
Conclusion

Design of a method to detect and precisely identify outliers in heterogeneous distributions:

- Structural and temporal similarity evaluation of distributions.
- Modelling of IP traffic as a link stream.

→ IP with anomalous degree profile, network scans.

Iterative removal of identified outliers

→ Validation: Creation of normal traffic (w.r.t $d_t(v)$).

⇒ Method applicable over temporal interactions in general.

Thank you for your attention!
**Identification: details**

**Detected Outlier** = time slice $T_i$ + degree class $C_j$

$$= \{(v, t) : v \in C_{41} \text{ and } t \in T_{504} = [1008, 1010]\}$$

Nodes $\in C_{41}$:

Abnormal on $I_{v_1} = [710.3, 713.1]$

Abnormal on $I_{v_2} = [628.8, 632.6]$

Abnormal on $I_{v_3} = [1007.3, 1010]$

Abnormal on $I_{v_4} = [440.5, 443.2]$

Identified outlier:

$$\{(v_3, t) : t \in I_{v_3}\}$$
Degree Profiles of 4 identified nodes

Audrey Wilmet, Tiphaine Viard, Matthieu Latapy, Robin Lamarche-Perrin
Degree Distribution: before and after

Fraction of \((v, t)\) s.t. \(d_t(v) = k\)

Fraction of \((v, t)\) s.t. \(d_t(v) > k\)

Degree \(k\)
Classes construction

Need to respect the heterogeneous nature of the distribution:
- have low degree couples \((v, t)\) which contains most of the traffic in isolated classes,
- take into account that the degree of nodes along time fluctuates and that generally: the larger the degree the larger the fluctuations.

\[\Rightarrow\text{logarithmic degree classes}\]

In logarithmic scale: points spaced of the same distance represent values in the same ratio \(r\)

\[
\begin{align*}
\text{lin: } & k_j \longrightarrow k_{j+1} = k_j + r \\
\text{log: } & k_j \longrightarrow k_{j+1} = k_j \times r \text{ and } \log(k_{j+1}) = \log(k_j) + \log(r)
\end{align*}
\]

\[
\begin{align*}
\text{In our method:} & \quad \log(k_{j+1}) = \log(k_j) + 0.1 \\
\text{Other construction:} & \quad \{1\}, \{2\}, ..., \{9\}, \{10, ..., 19\}, \{20, ..., 29\}, \ldots, \{90, ..., 99\}, \{100, ..., 199\}, \text{ etc.}
\end{align*}
\]