# Outliers Detection using the Instantaneous Degree within IP Traffic Modelled as a Stream Graph

Audrey Wilmet[1], Matthieu Latapy[1],
Robin Lamarche-Perrin[2]

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

RESCOM 2018, Toulouse, January 18, 2018



# Context and Goals

- IP traffic:

Flow of data across the Internet:

→ nodes: IP addresses;

→ interactions: packet exchanges;

→ temporal interactions.

- Goals:

① Detect outliers = irregular patterns:

→ global high activity of an IP address;

→ sharp variation in an the activity of an IP address;

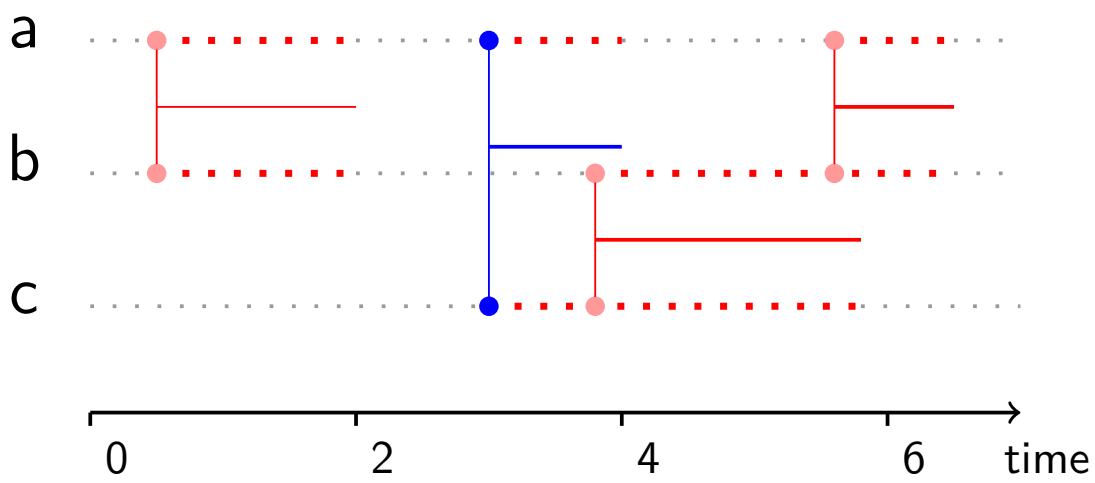
→ high activity at a particular moment.

② Identify them: find their exact cause in the data;

③ remove them precisely from the traffic.

# IP Traffic as a Stream Graph

Stream graph: 1h of real IP Traffic (MAWI)




*ex:* $a$ and $c$ interacted from $t_1 = 3$ to $t_2 = 4$

→ M. Latapy *et al.*, 2017 ; T. Viard *et al.*, 2017.

- Other Data Modelling: Signal, graph

→ Y. Himura *et al.*, 2013 ; H. Asai *et al.*, 2014.

# Our Approach

Data Modelling: temporal interactions = stream graph

---

Outlier = entity statistically deviating from others

1) **Entity** $x$: $(uv, t)$ $uv$ $(v, t)$ $v$ $t$ $\Rightarrow (v, t)$

2) **Studied set** $x \in X$:

All entities of the same types or a subset of it ?

$$\Downarrow \qquad\qquad \Downarrow$$

$$t \in X = T \qquad\qquad t \in X \subset T$$

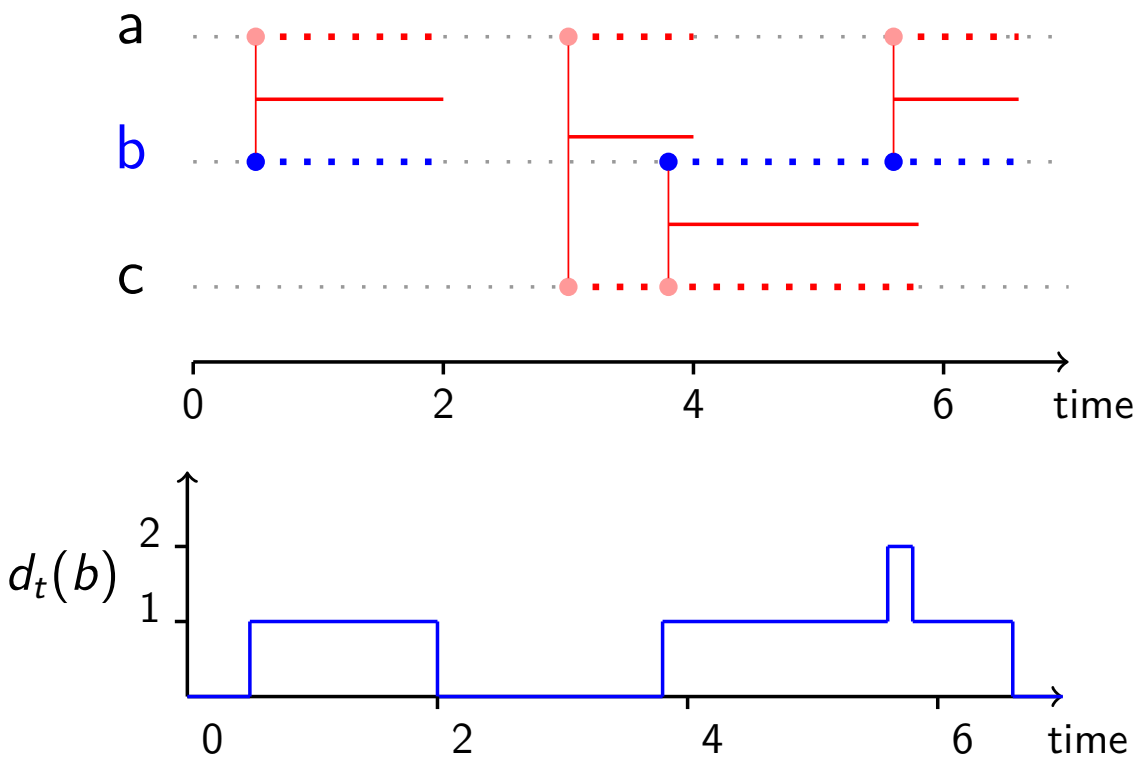
3) **Feature** $\mathcal{F}$ on $X$ $\Rightarrow$ Instantaneous degree

# Instantaneous Degree of $(v, t)$

Number of neighbours of node $v$ at time $t$:

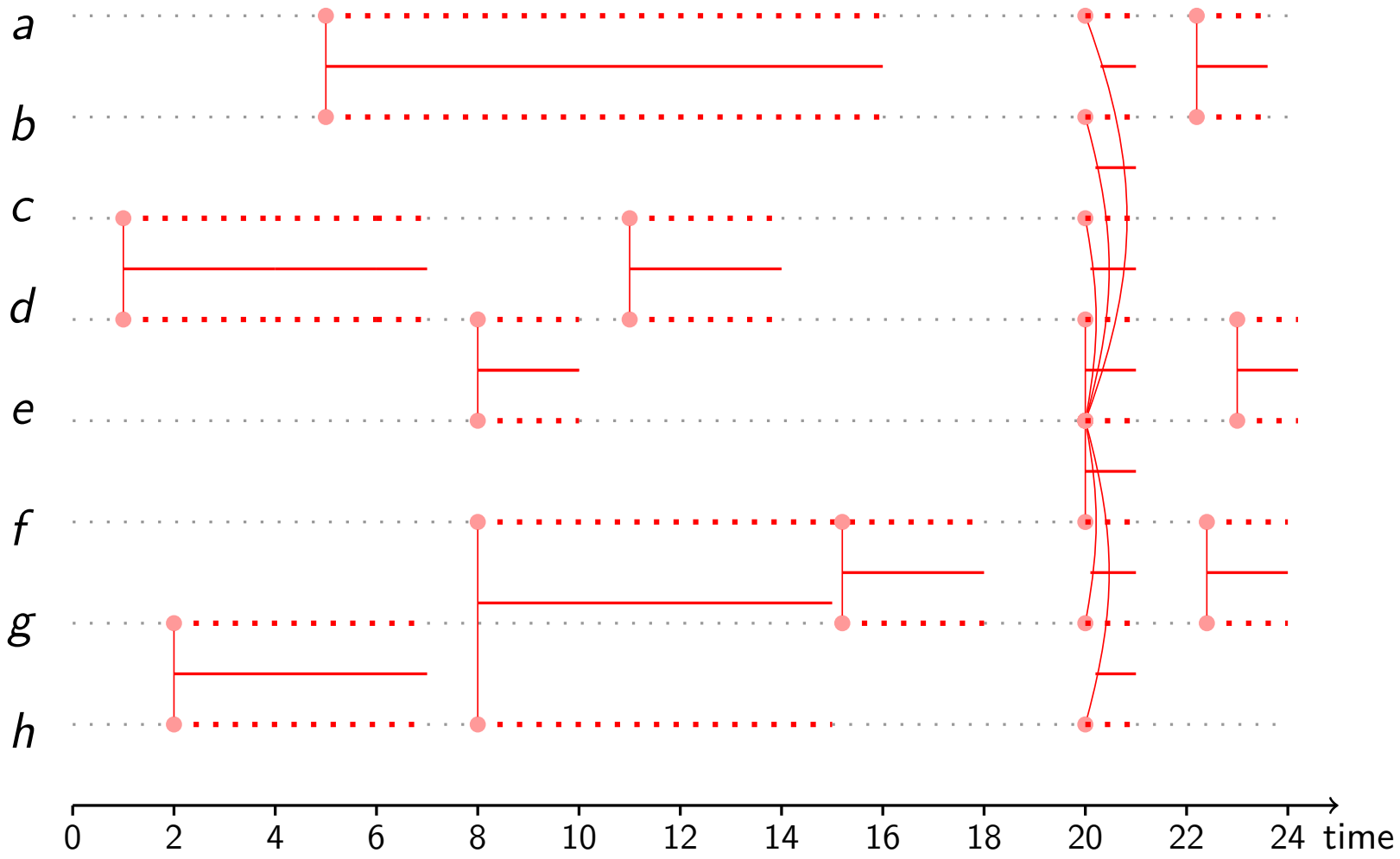
$$d_t(v) = |\{u, (t, uv) \in E\}|$$

*Example: degree profile of b*



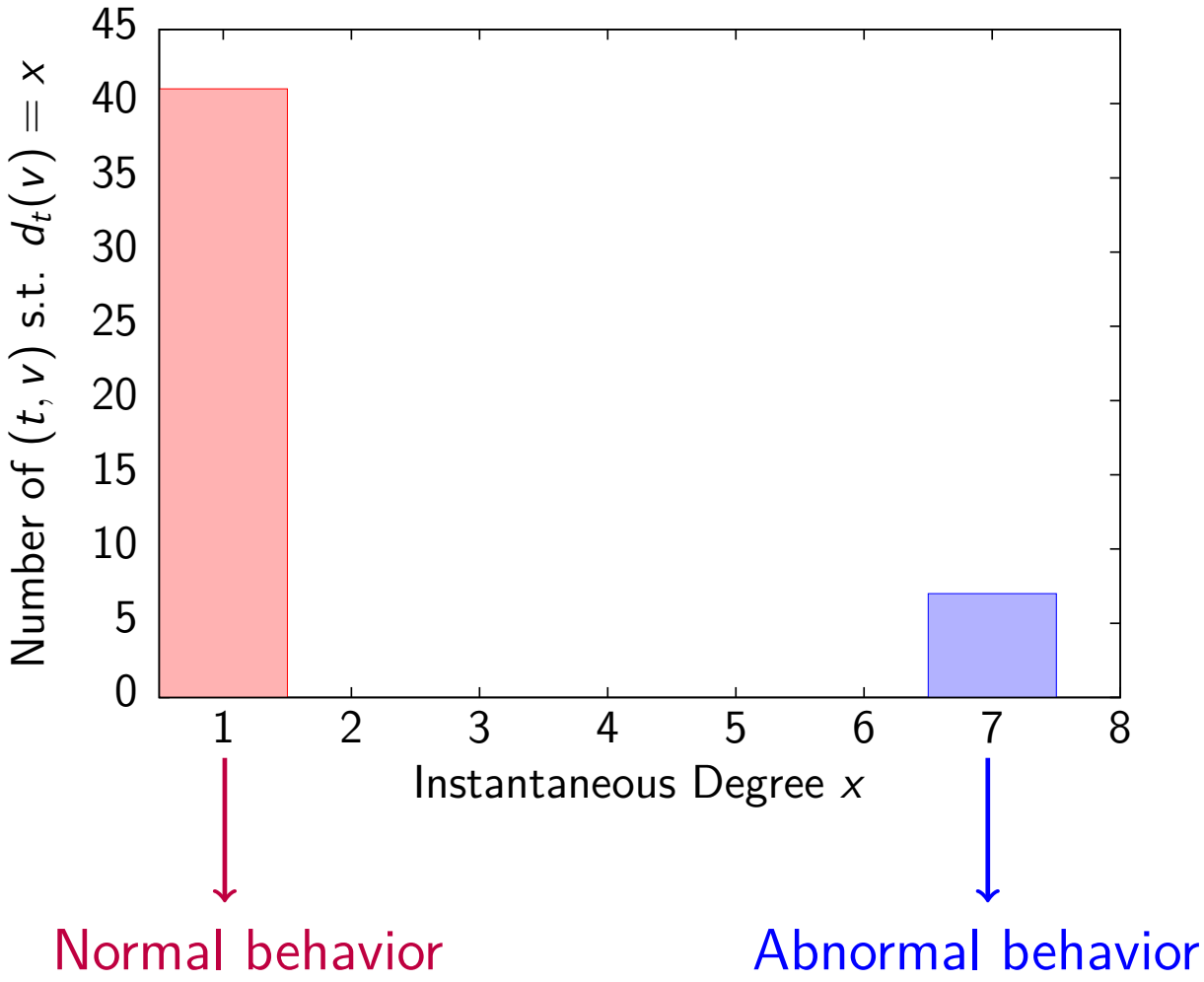
# ① Detection

**Entity:** $(v, t)$ **Feature:** $d_t(v)$ **Studied set:** $V \times T$



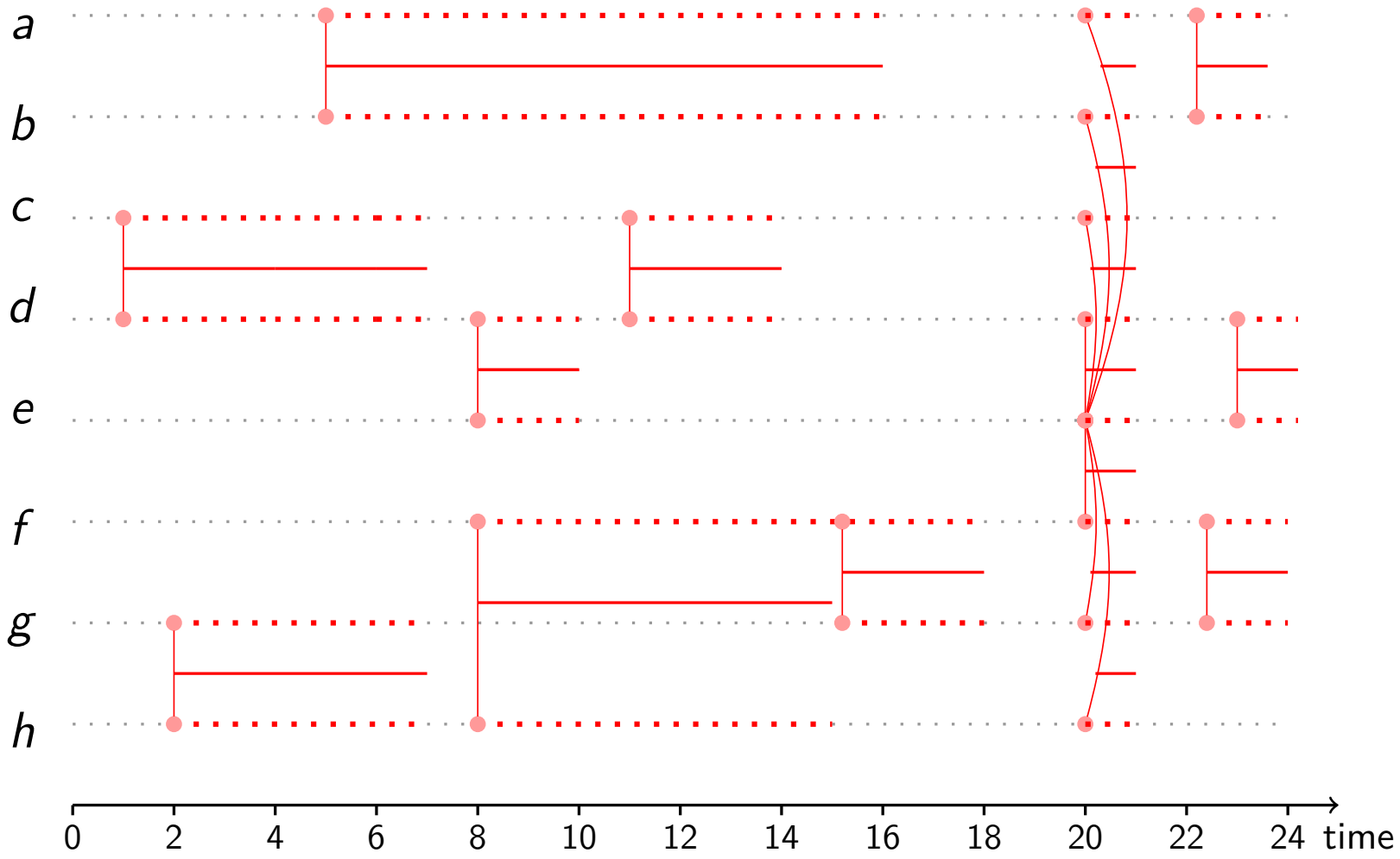
# ① Detection

**Entity:** $(v, t)$ **Feature:** $d_t(v)$ **Studied set:** $V \times T$



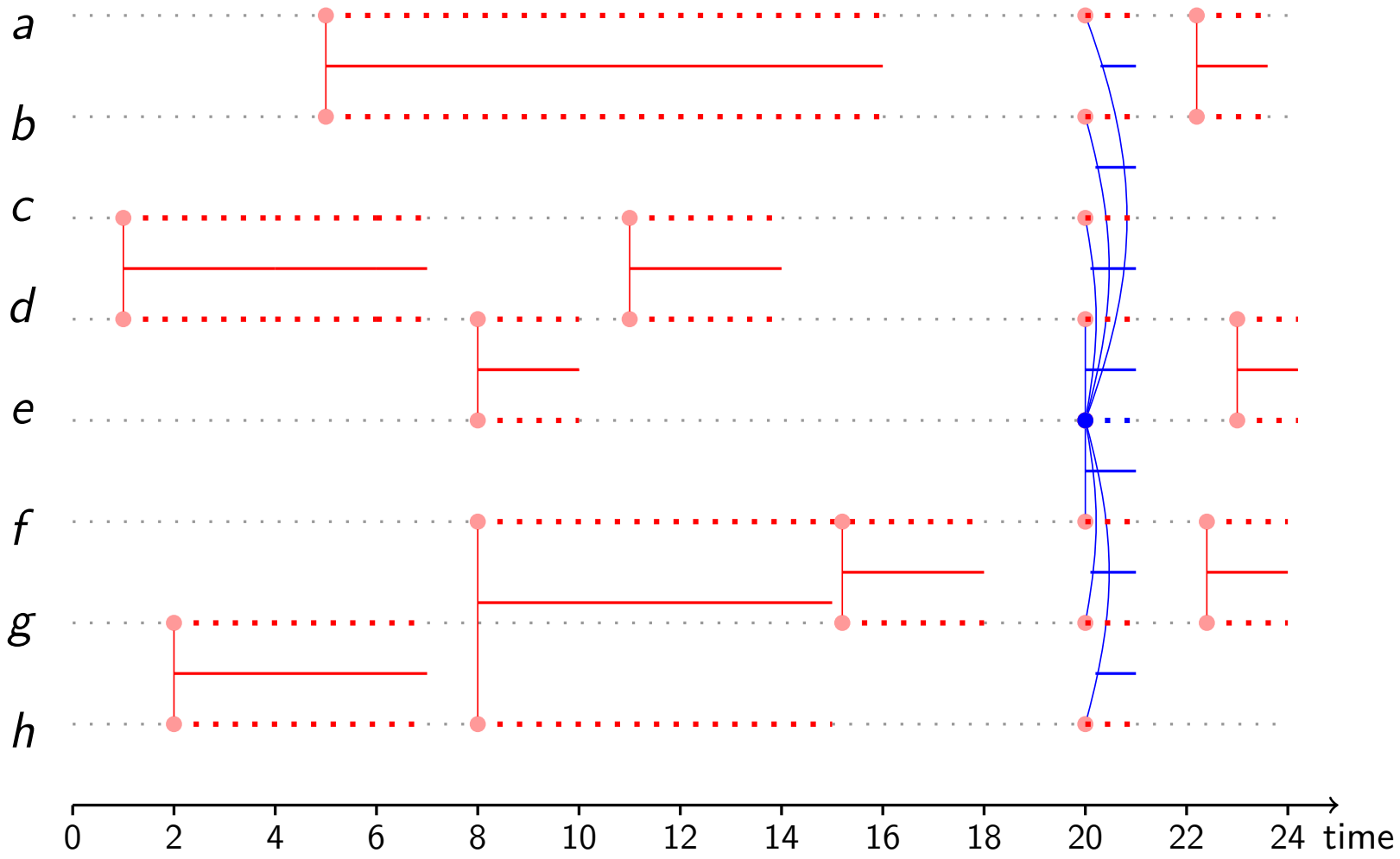
# ① Detection

$$\{(v,t) : d_t(v) = 7\} \quad \Rightarrow \text{Detected outlier}$$



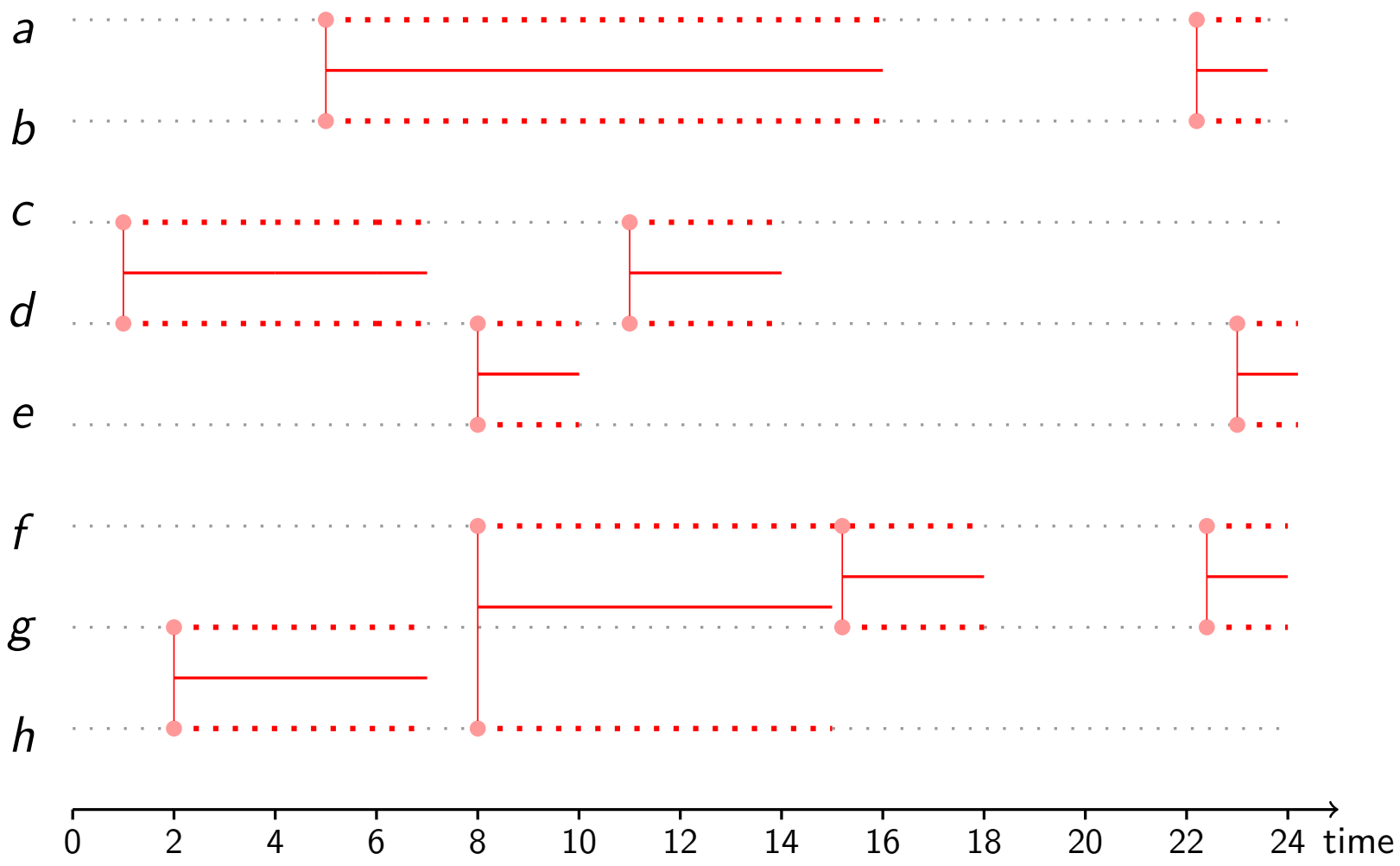
# ② Identification

$\{(e, t), t \in [20, 21[\} \quad \Rightarrow$ identified outlier



# ③ Removal

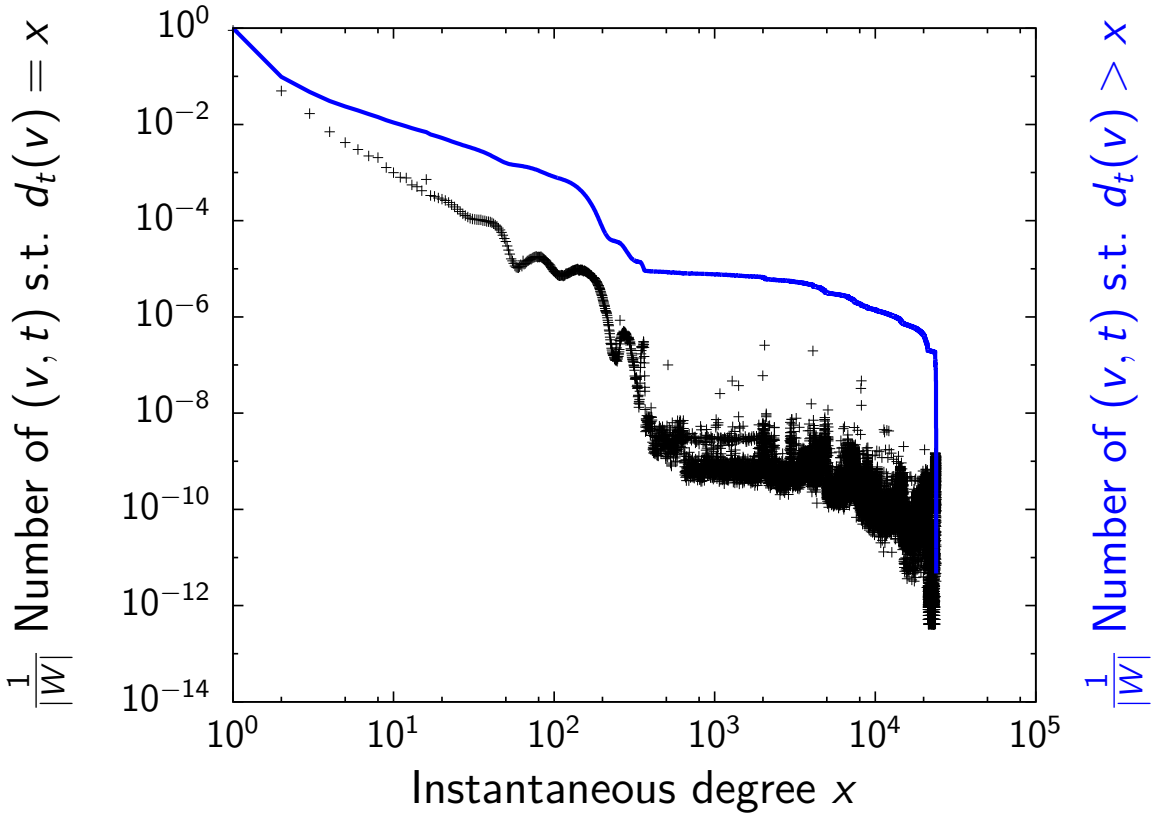
$\{(e, t), t \in [20, 21[\}$



# In Practice

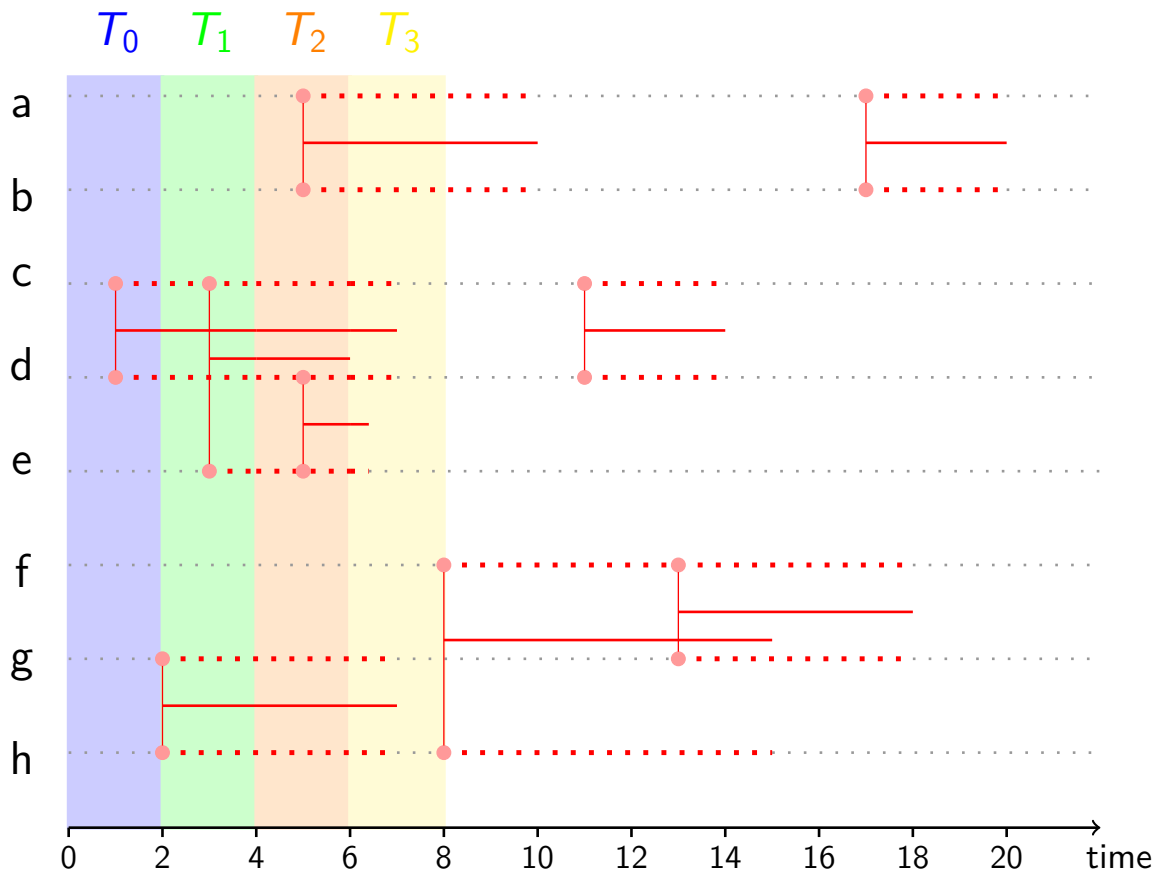
# Global Degree Distribution

**Entity:** $(v, t)$
**Feature:** $d_t(v)$
**Studied Set:** Global, $W = V \times T$.

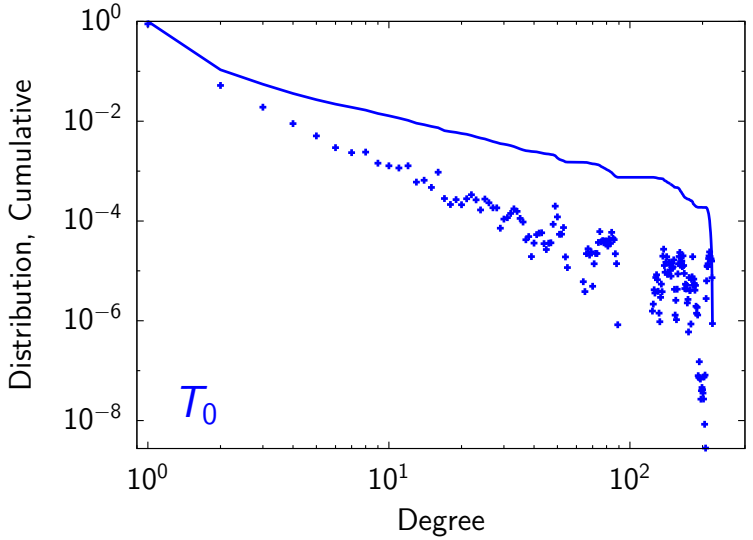


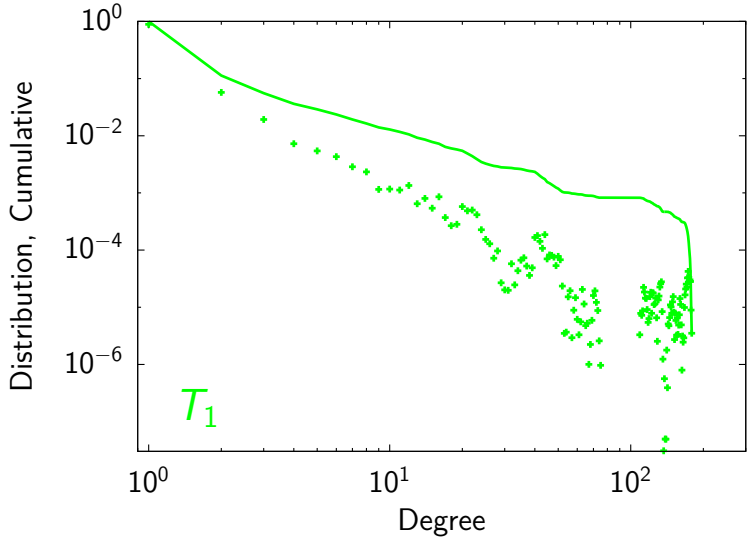
# Local Degree Distribution (1/2)

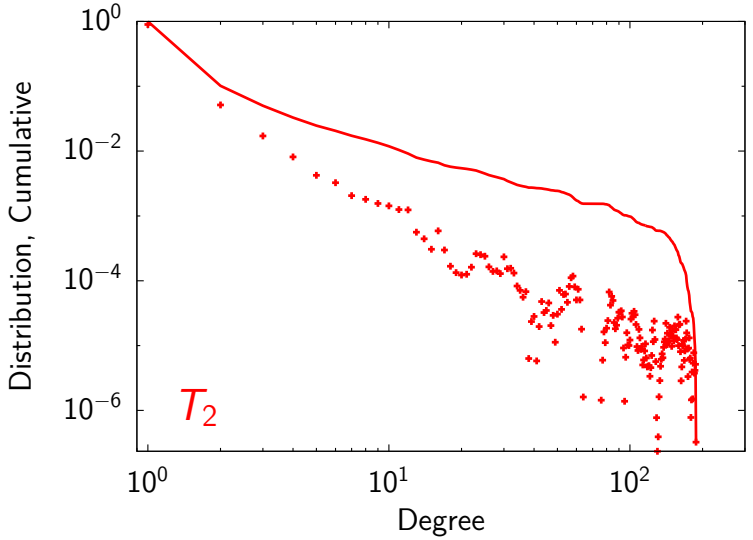
**Entity:** $(v, t)$
**Feature:** $d_t(v)$
**Studied Set:** Local, $W_i = T_i \times V$, $T_i = [2i, 2i + 2[$.

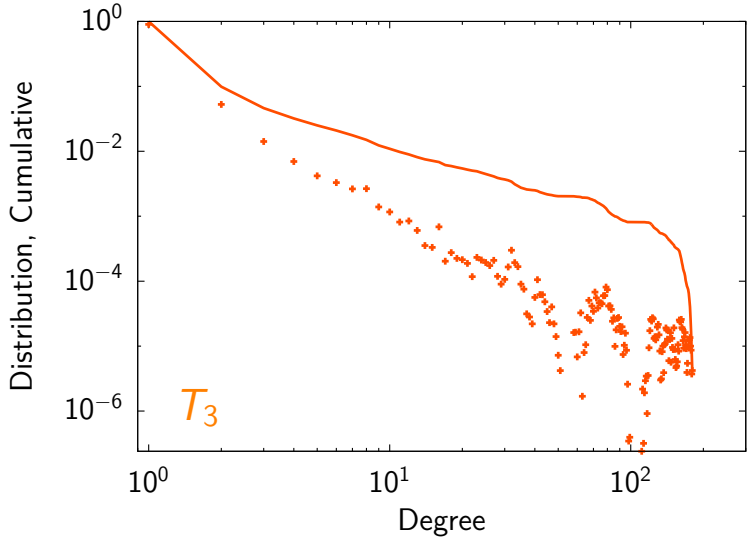


# Local Heterogeneity (2/2)



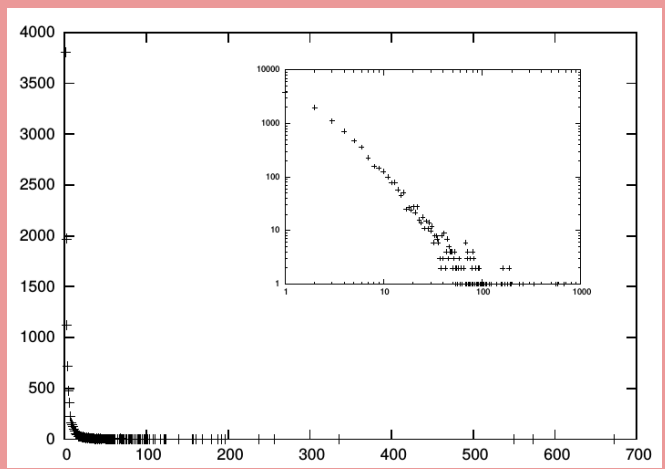



# Difficulties

Outlier = Activity that deviates from the usual one

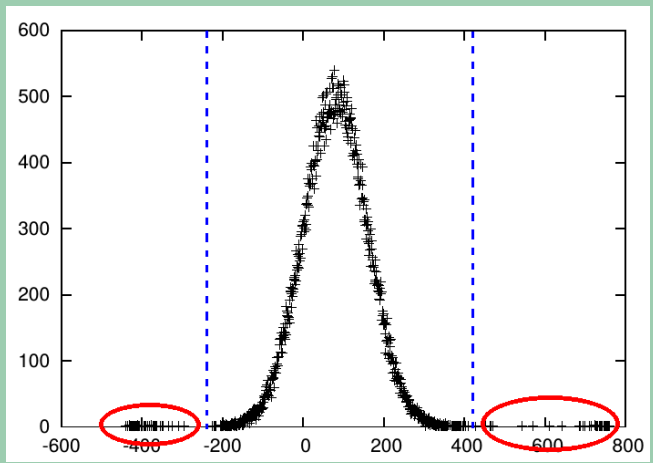
Find an outlier $\Longleftrightarrow$ Find the normality

✗




Heterogeneous

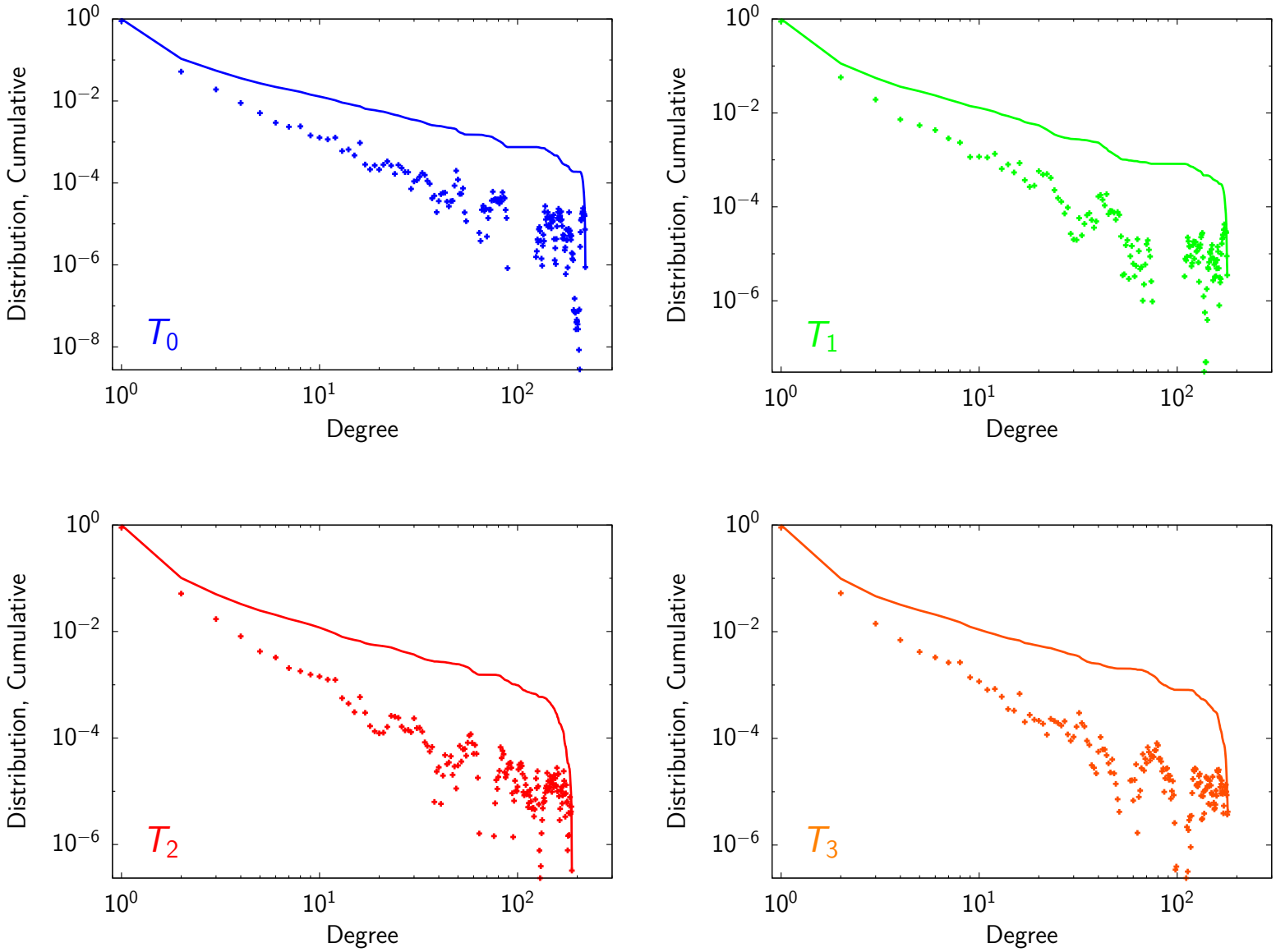
✓



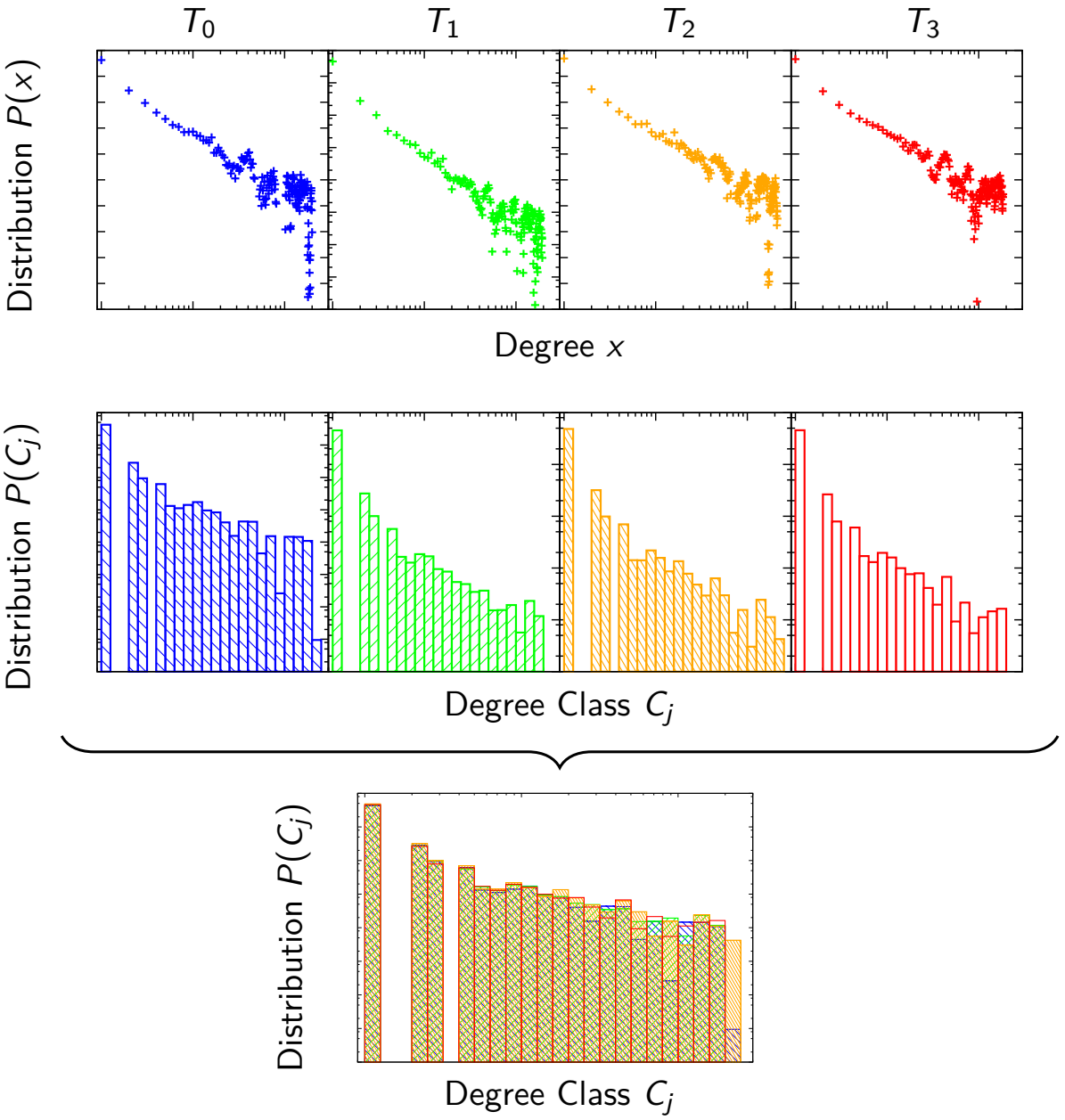

Homogeneous with outliers

# Our Method

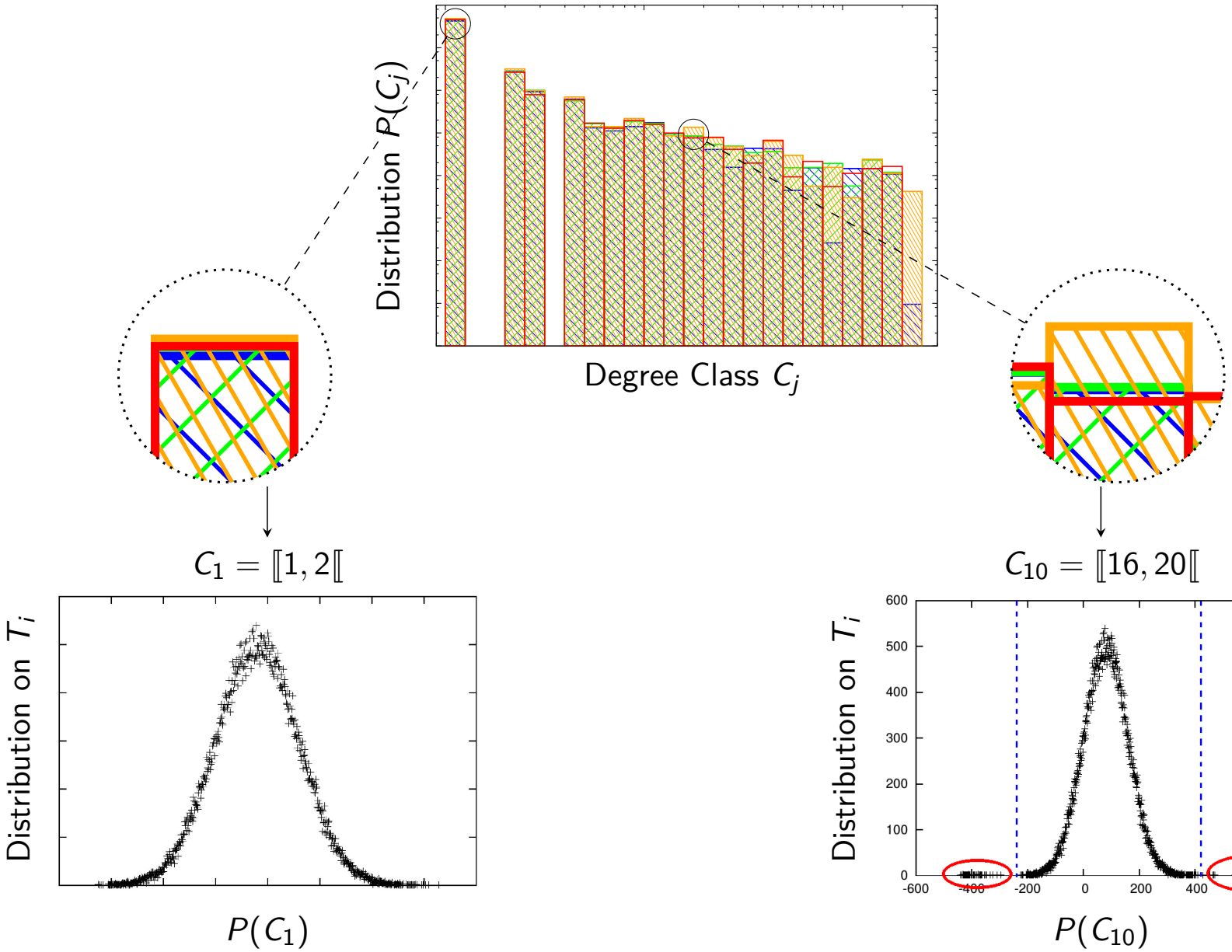
# Temporal Homogeneity



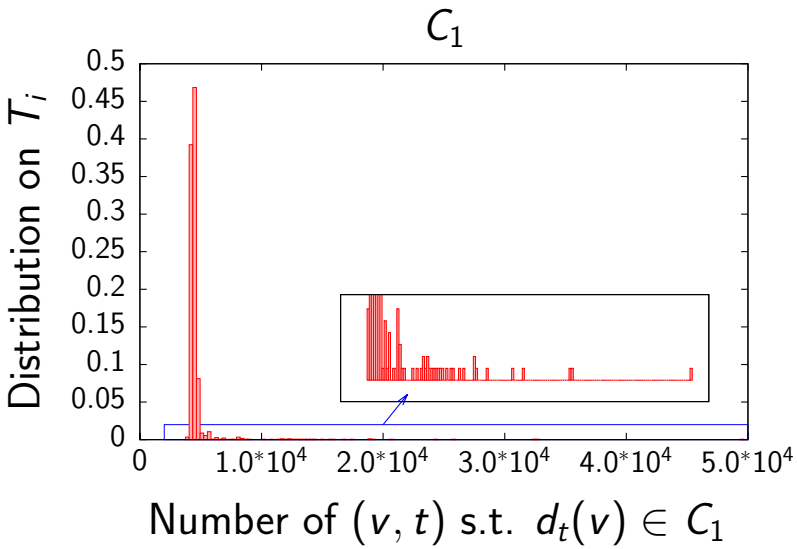
# Comparison of Local Distributions

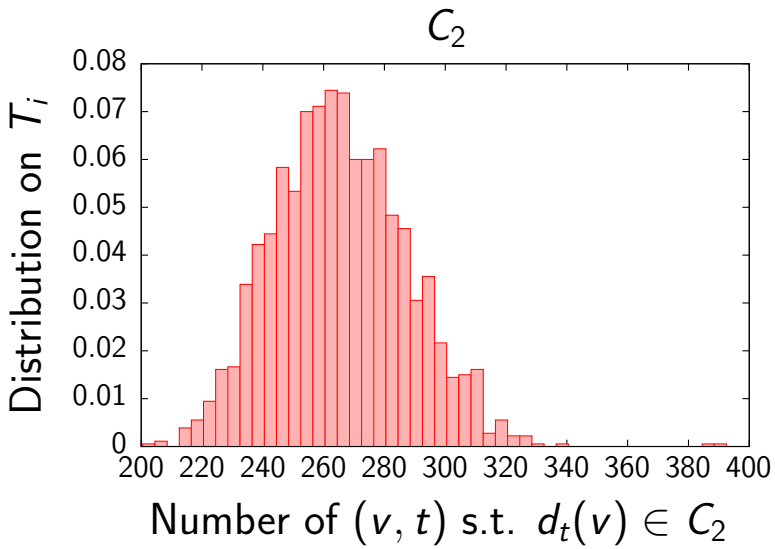


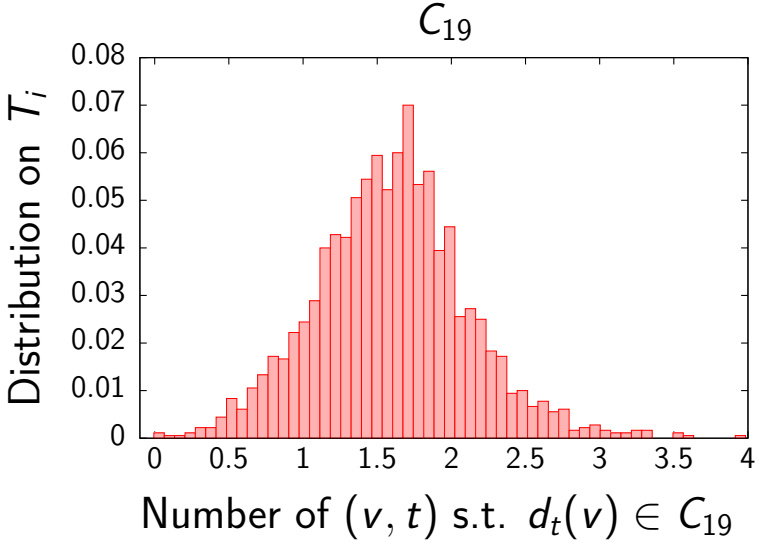
# Comparison of Local Distributions

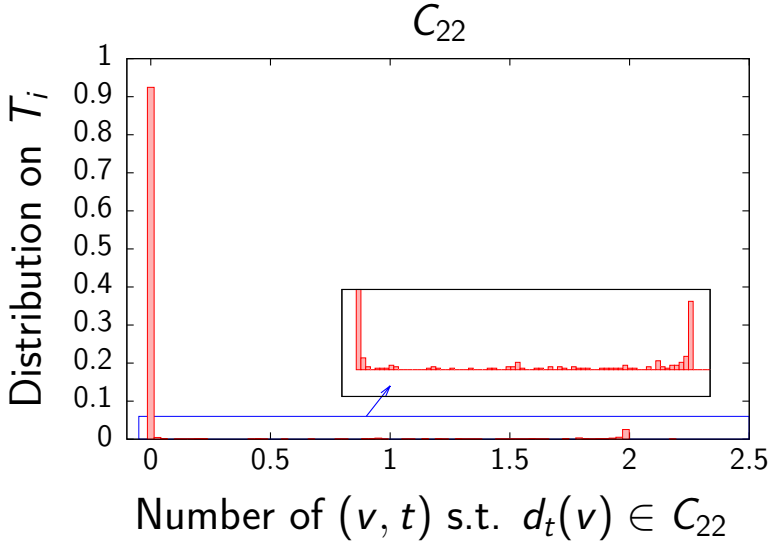


# Results: Homogeneous Distributions

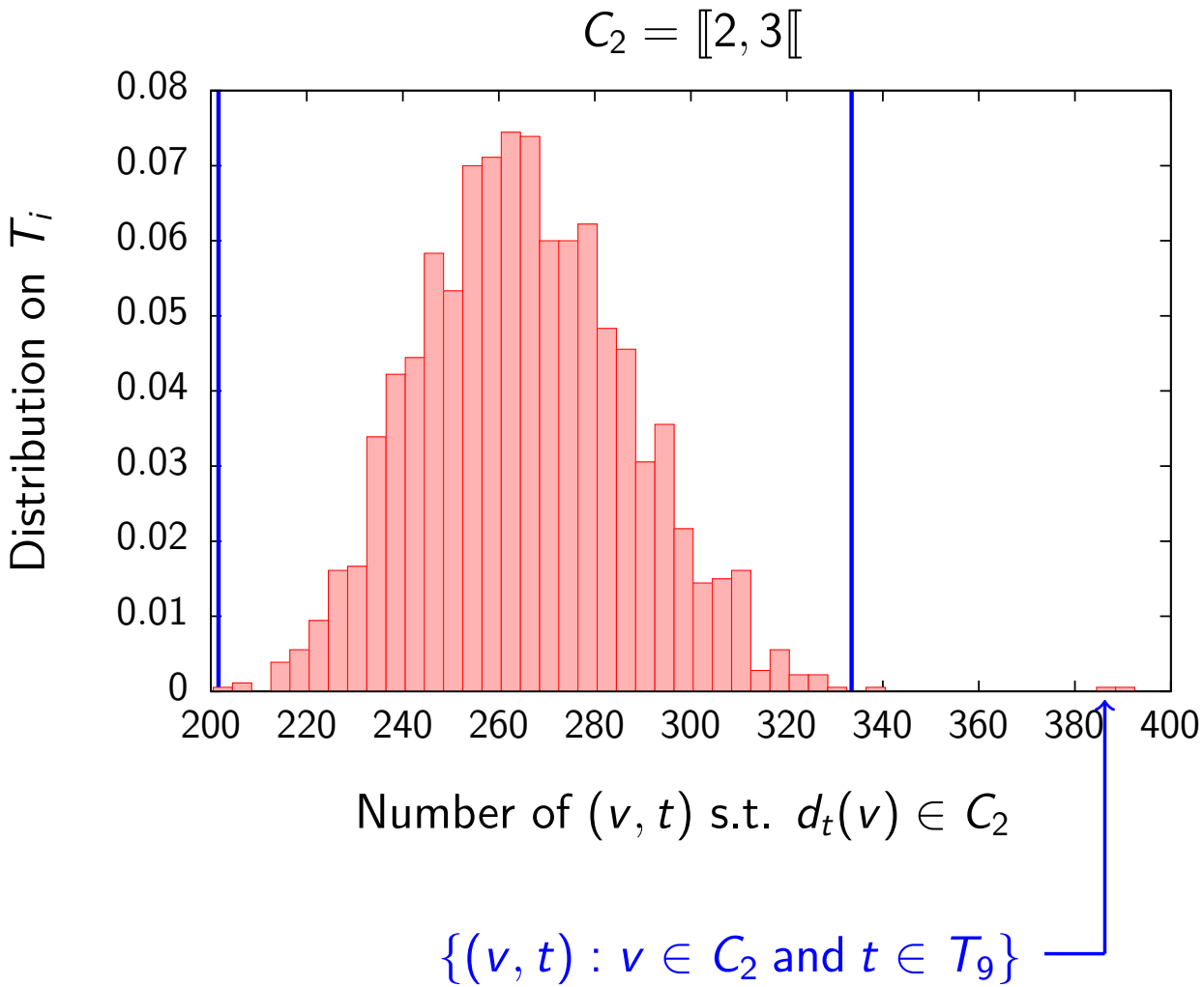


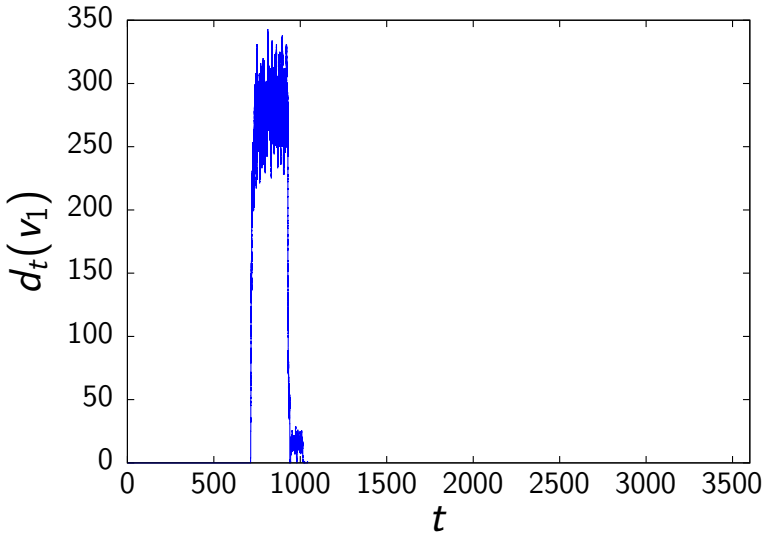
# Results: Homogeneous Distributions

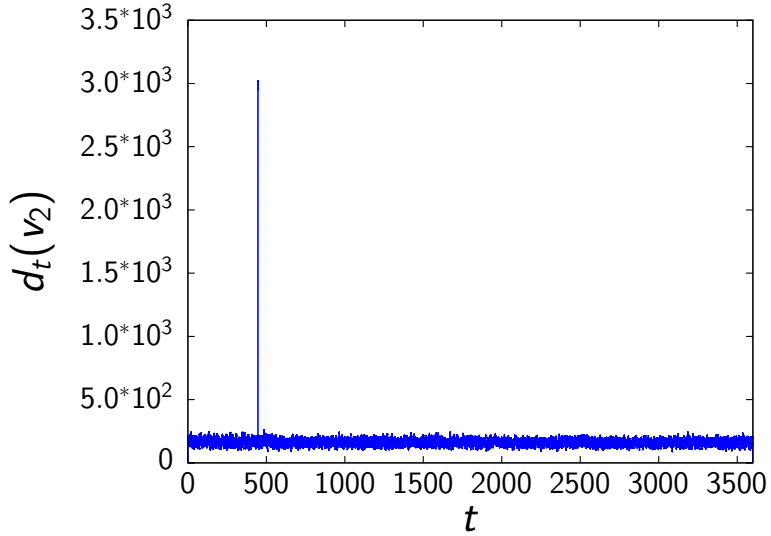


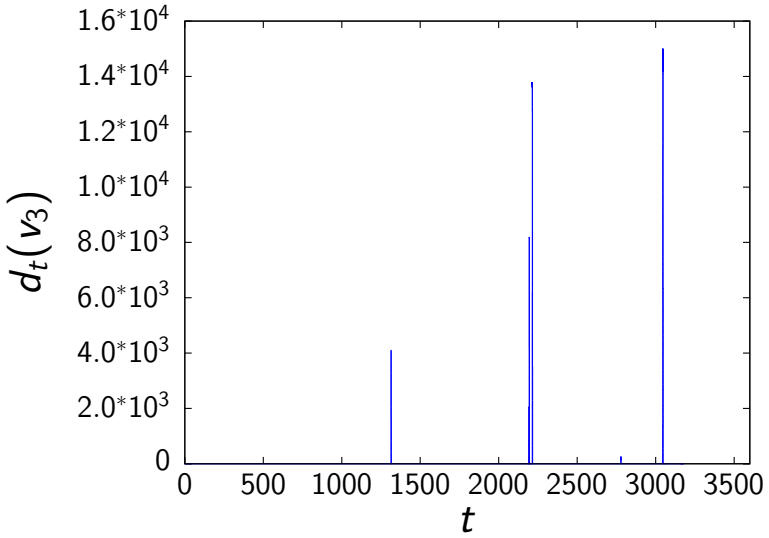

Outlier = temporal $T_i$ + structural $C_j$ information.

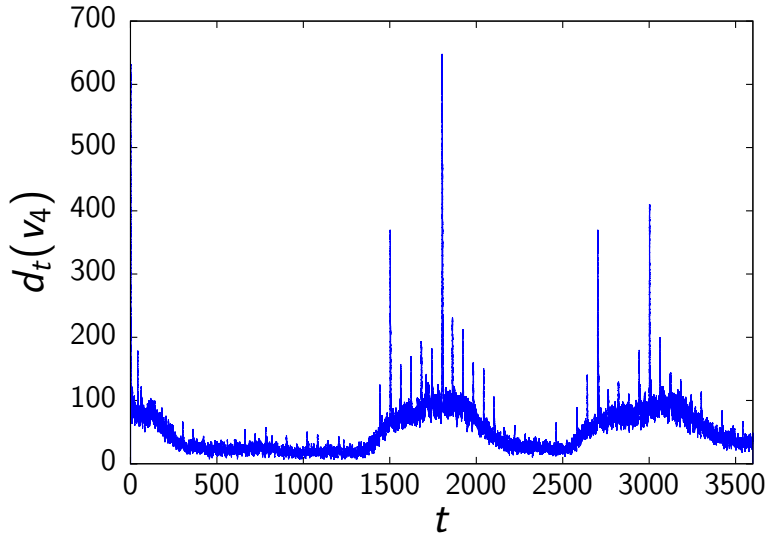
# Identification and Removal - Degree Profiles

$\Longrightarrow$ Suspicious couples $(v, T_i)$.

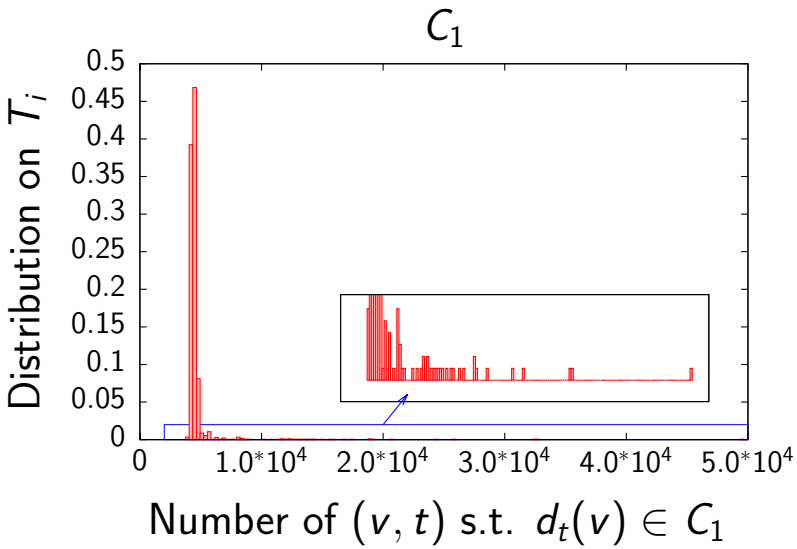


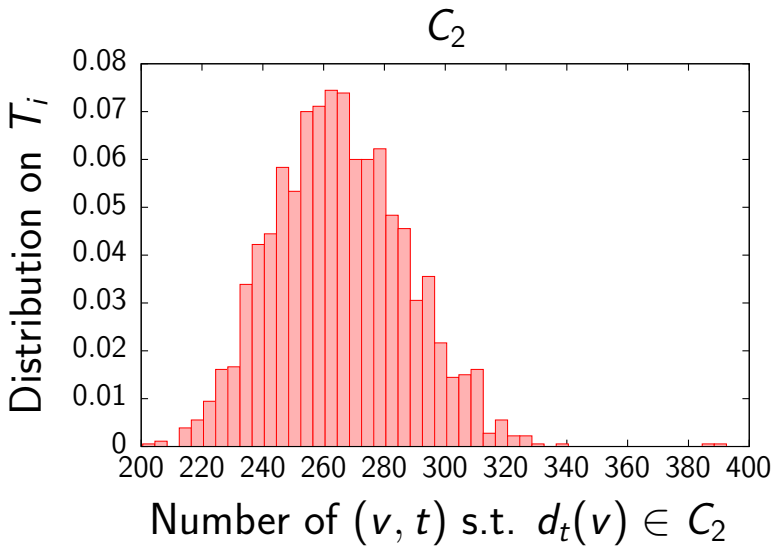



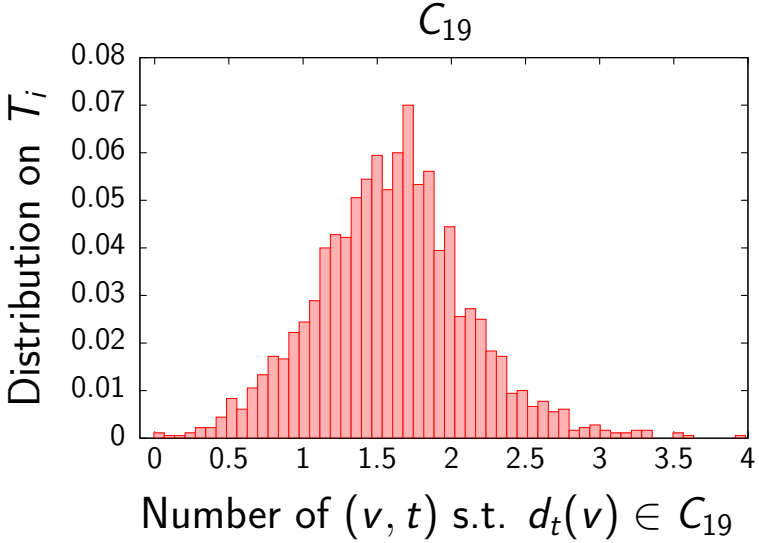



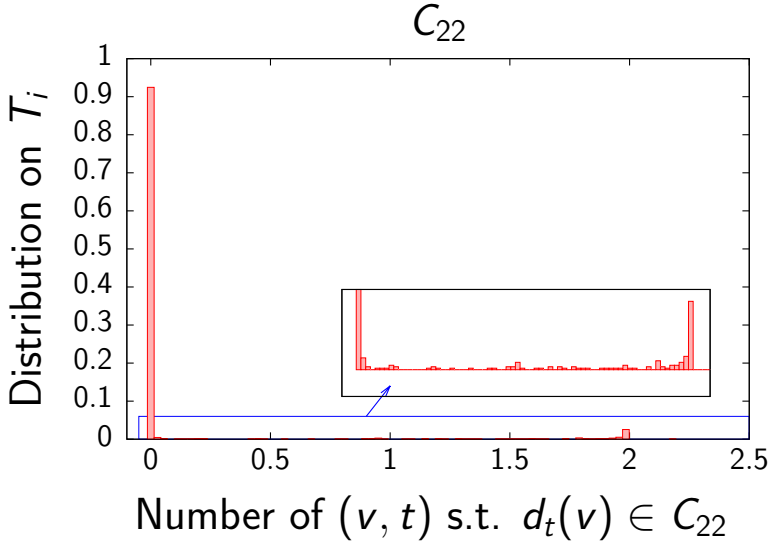



# Distributions after removals

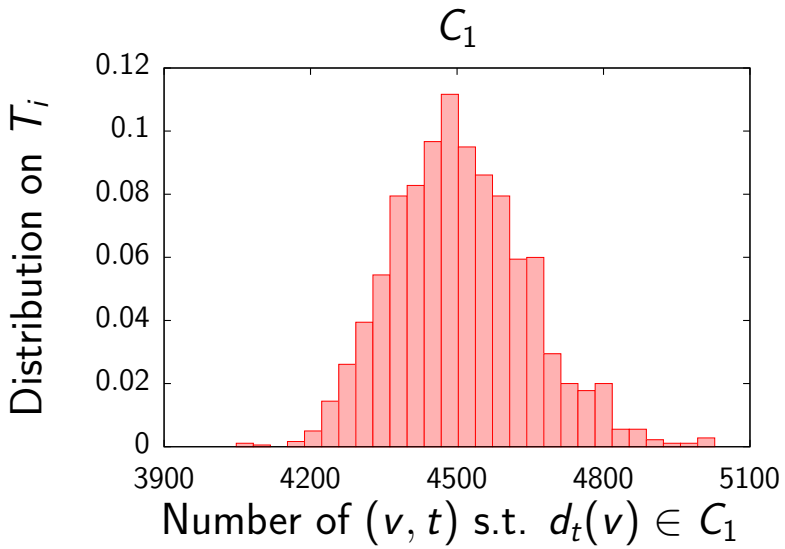


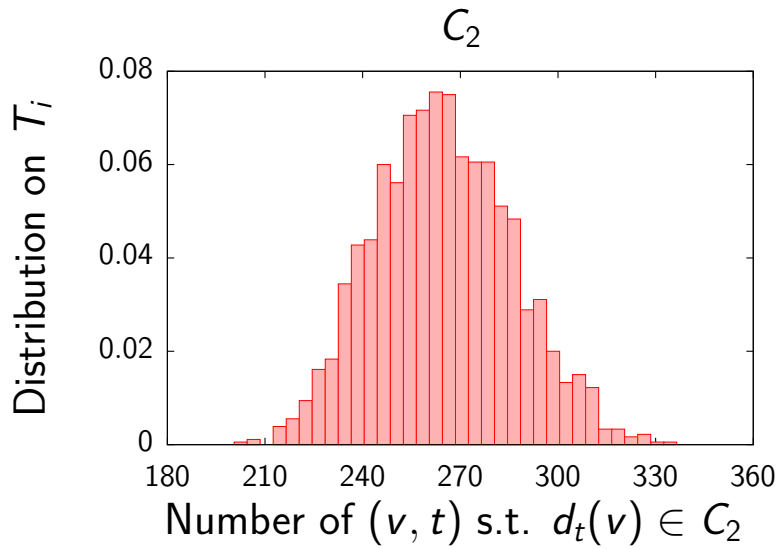



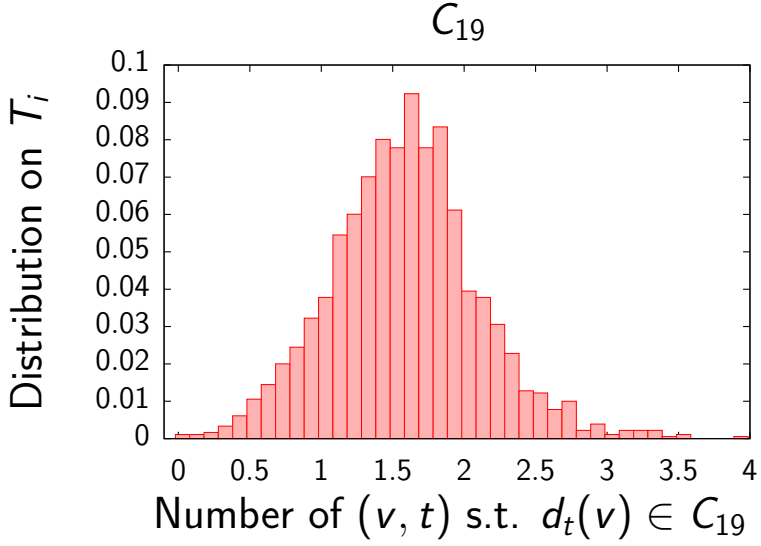



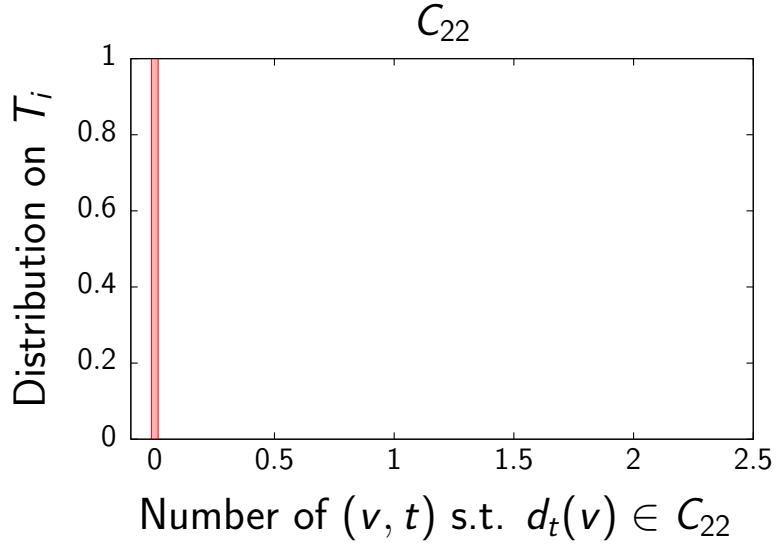



# Distributions after removals



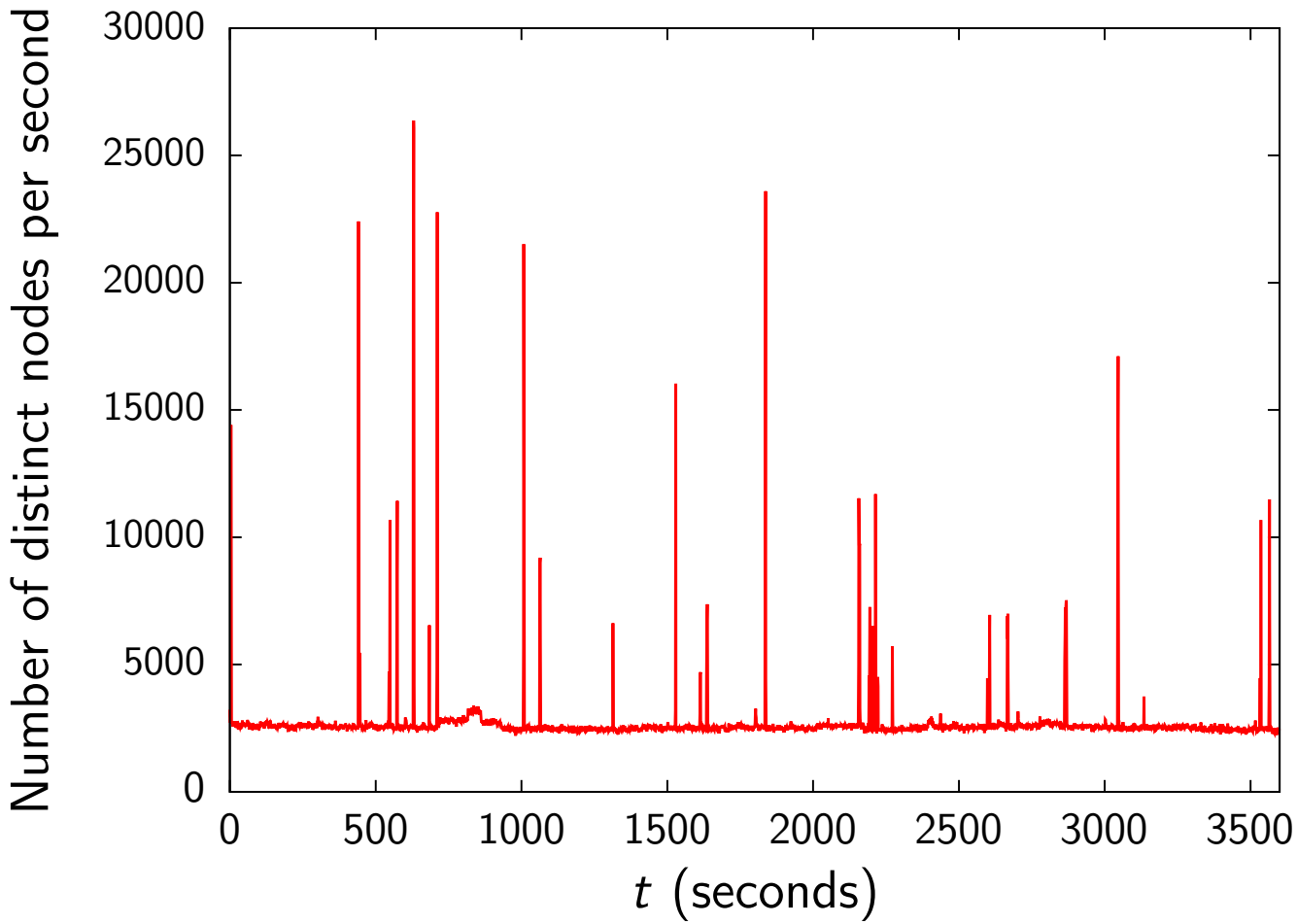



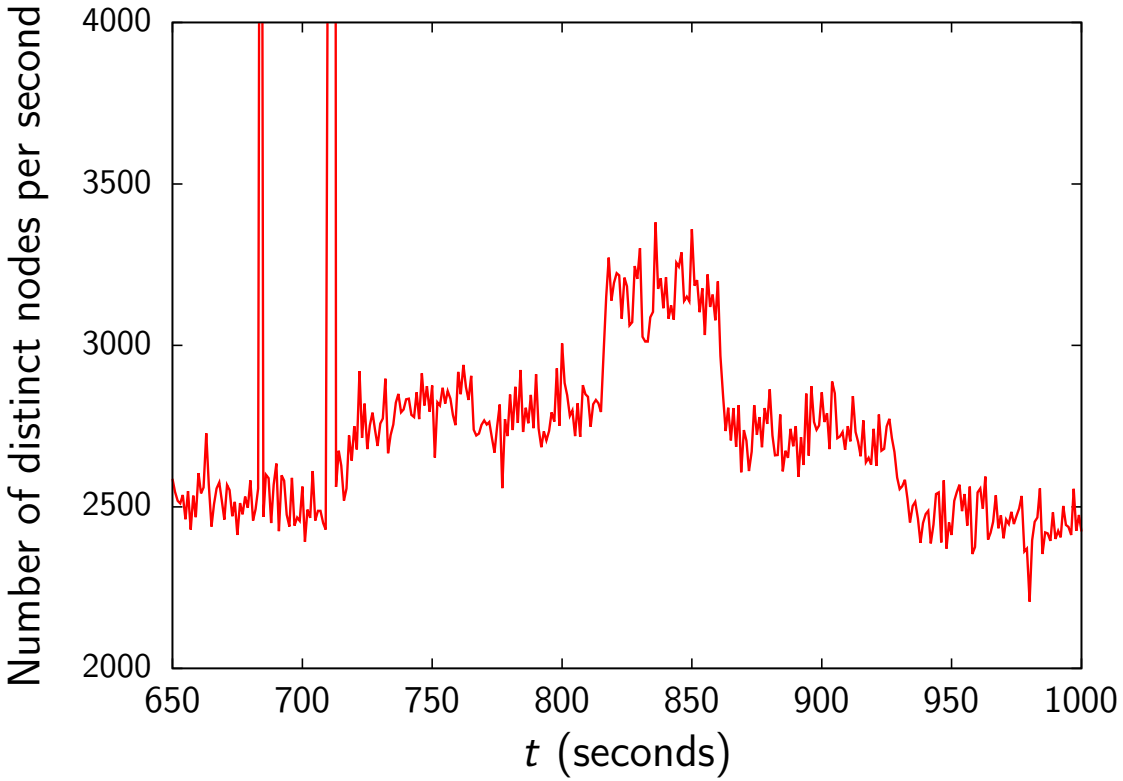
# Removal of identified outliers $(v, T_i)$

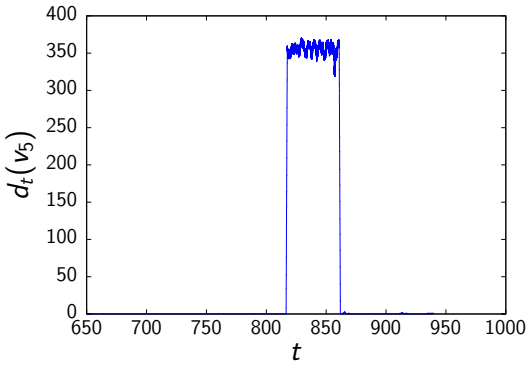
$\Longrightarrow$ Consequence on the number of distinct nodes per second.

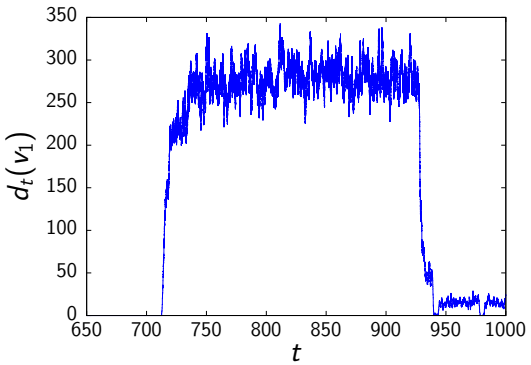


# Removal of identified outliers $(v, T_i)$

$\Longrightarrow$ Consequence on the number of distinct nodes per second.

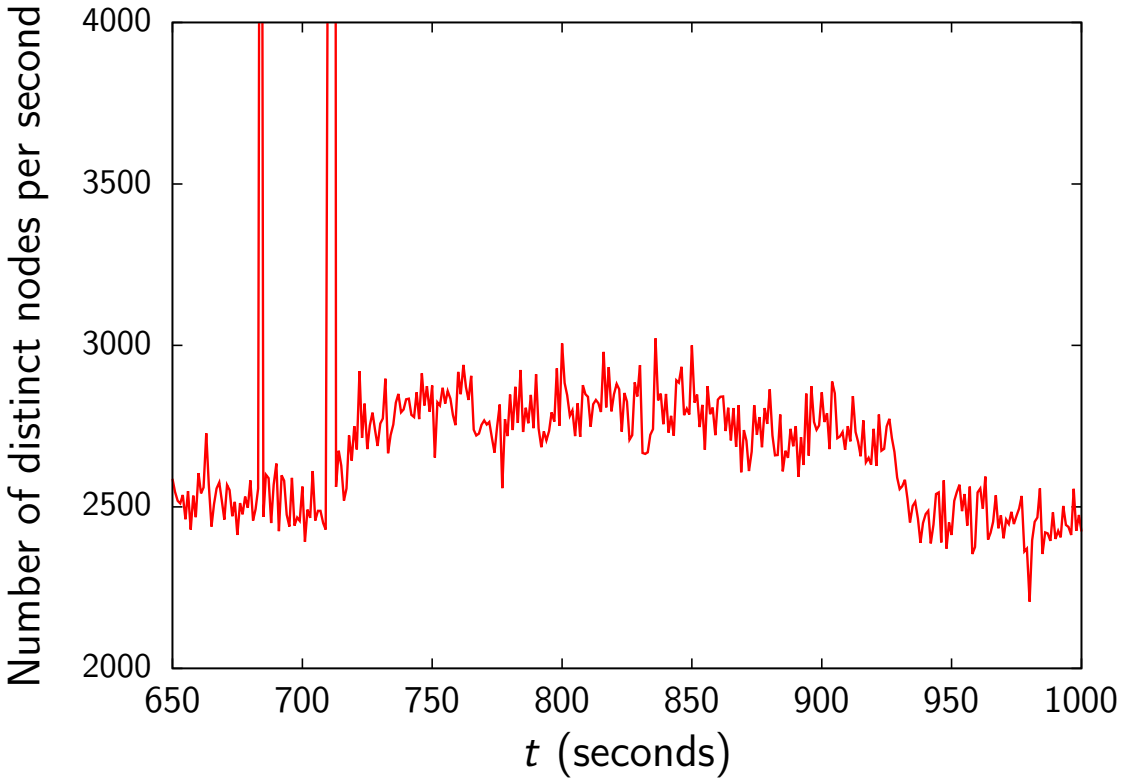


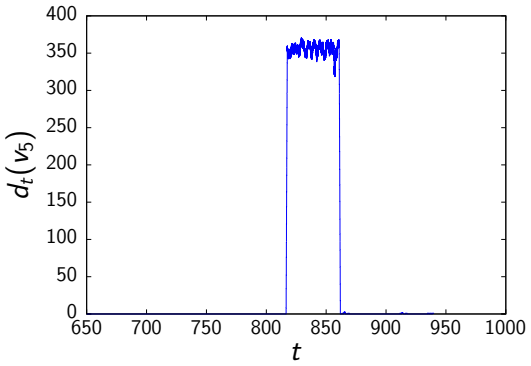



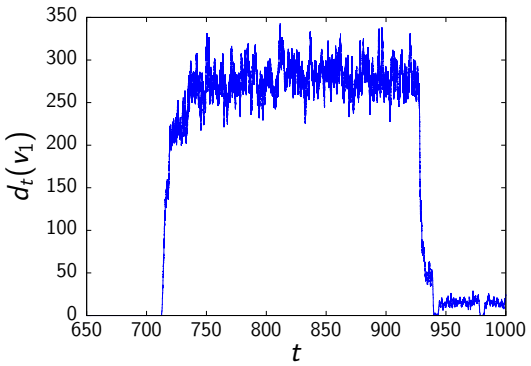



# Removal of identified outliers $(v, T_i)$

$\Longrightarrow$ Consequence on the number of distinct nodes per second.

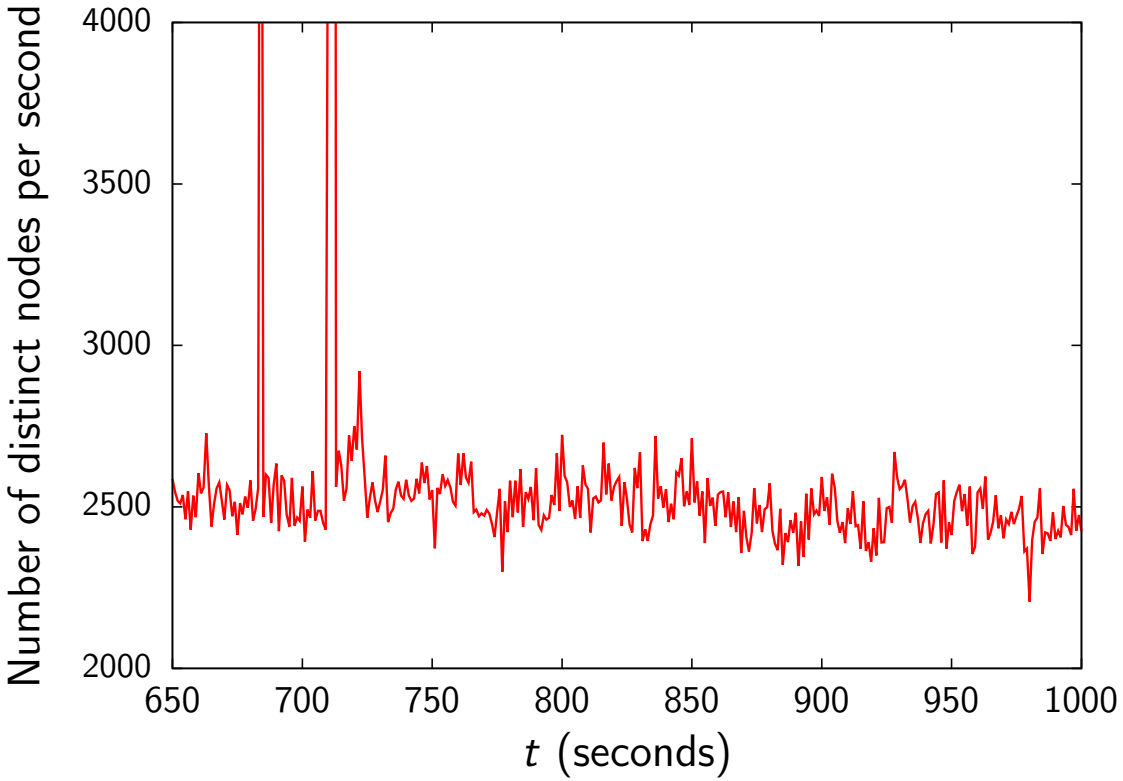


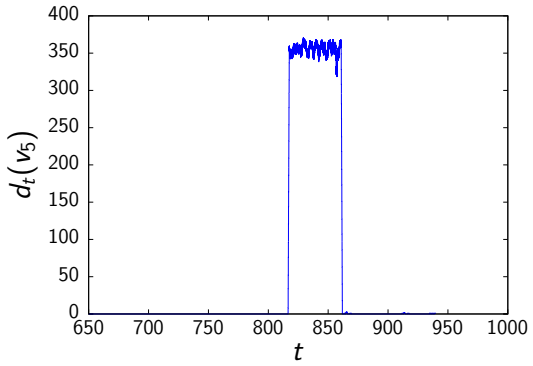



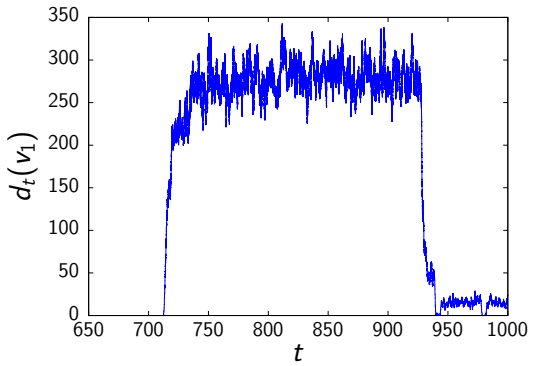



# Removal of identified outliers $(v, T_i)$

$\Longrightarrow$ Consequence on the number of distinct nodes per second.

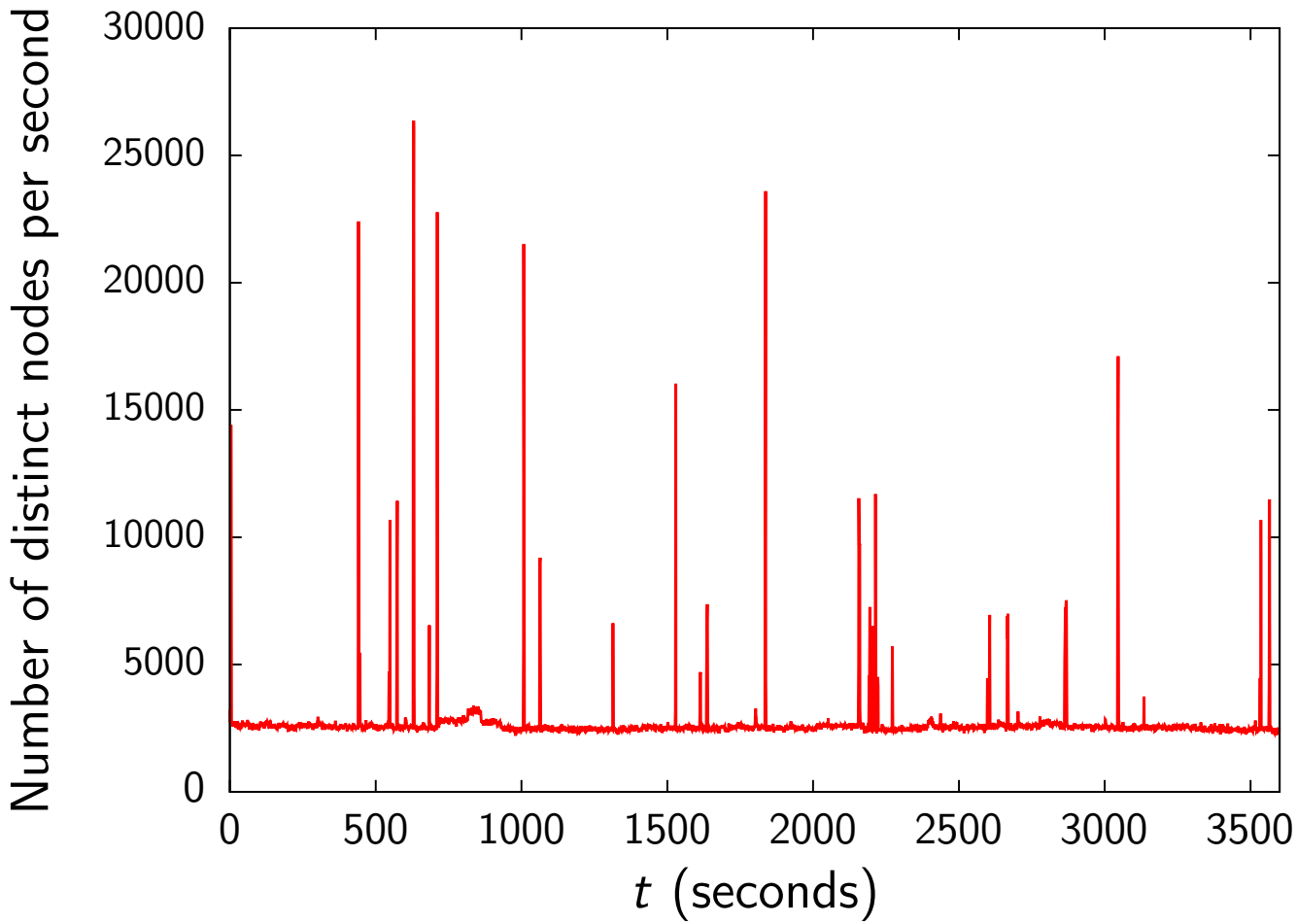


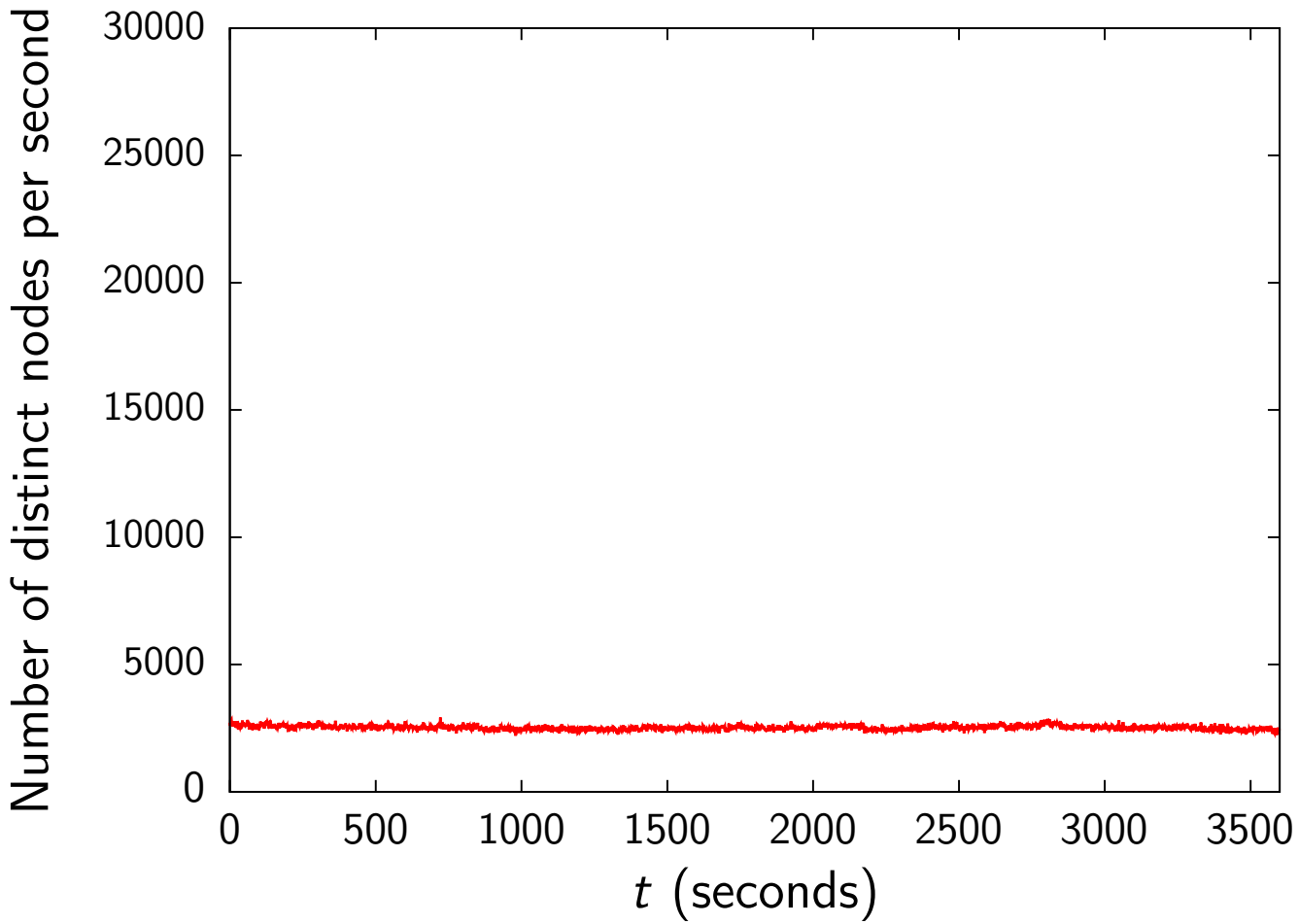
# Removal of identified outliers $(v, T_i)$

$\Longrightarrow$ Consequence on the number of distinct nodes per second.



# Removal of identified outliers $(v, T_i)$

$\Longrightarrow$ Consequence on the number of distinct nodes per second.



# Conclusion

# Conclusion

- Contributions:

– Modelling of IP traffic as a stream graph

– Design of a method to detect outliers in heterogeneous distributions

→ IP with anomalous degree profile, network scans.

– Iterative removal of identified outliers

→ Return to normal traffic (w.r.t $d_t(v)$).

⇒ Method applicable over temporal interactions in general.

- Several possible improvements:

– Identification of couples $(v, t)$ instead of couples $(v, T_i)$,

– more complex feature: clustering coefficient,

– exploring other assumptions:

→ T. Schieber *et al.*, Q. Zhang *et al.*, 2017.

# Thanks for your attention !