Temporal Graph Mining for Fraud Detection
Part III

Christos Faloutsos
CMU

Pedro Fidalgo
Mobileum

Mirela Cazzolato
USP

Bird’s eye view

- Part#1: Introduction – types of fraud

- Part#2: Graphs Mining – patterns and tools

- Part#3: Visualization - conclusions

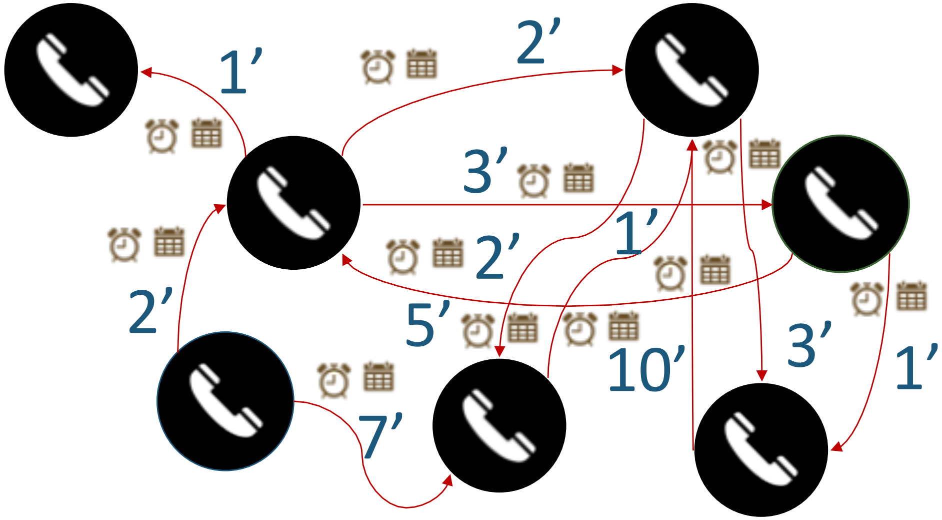
‘Recipe’ Structure:

- **Problem definition**
- Short answer/solution
- LONG answer – details
- Conclusion/short-answer

Problem definition

Given: who-calls-whom, when, and for how long network

Real life:
<u>millions</u> of
calls per day



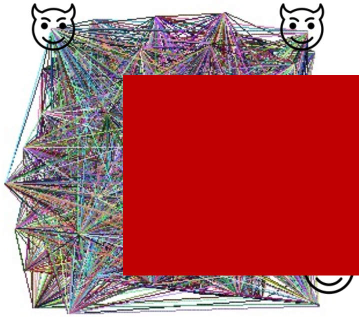
Costumers

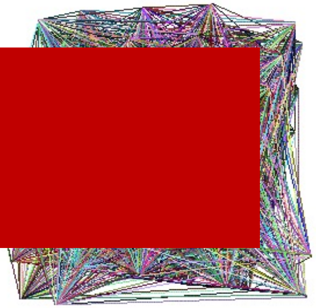
Fraudsters

Find: nodes with <u>strange behavior</u>

Problem definition

(source, destination, timestamp, duration)





Find fraudsters and explain why

Case 1:

(semi-) supervised:

some labels

Case 2:

Un-supervised:

no labels

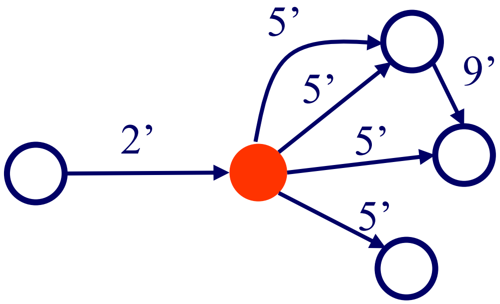
‘Recipe’ Structure:

- Problem definition
- **Short answer/solution**
- LONG answer – details
- Conclusion/short-answer

How to proceed?

One approach

- Extract features from each node
  - thus, *n*-d vectors
- look for anomalies
- and plot



‘Recipe’ Structure:

- Problem definition
- Short answer/solution
- **LONG answer – details**
- Conclusion/short-answer

Bird’s eye view

- …
- 4. Time evolving graphs
- 5. Visualization - practitioner’s guide
  - Which features?
  - Outlier detection?
  - Visualization tools?
  - Case studies
- 6. Conclusions

Which features?

Which features?

A: ones that spot known types of fraud:

- **‘brushing’:**
- **telemarketers:**
- **Wangiri:**
- **DDoS:**
- **Lockstep / collusion:**

Cazzolato, M.T., Vijayakumar, S.; Lee, MC.; Vajiac, C.; Park, N.; Fidalgo, P.; Traina, A.J.M.; Faloutsos, C., *CallMine: Fraud Detectoin and Visualization of Million-Scale Call Graphs*. ACM CIKM, 2023.

Which features?

from part I

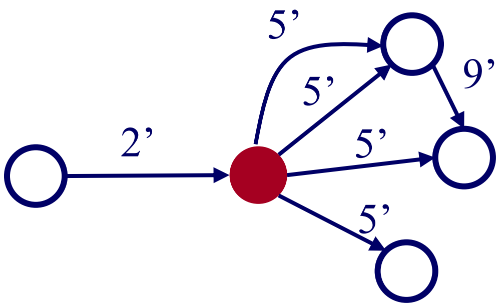
A: ones that spot known types of fraud:

- **(F1) 'brushing' :** degree; velocity
- **(F3) telemarketers:** out-degree; inter-arrival times
- **(F3) Wangiri:** duration (median; variance)
- **(F3) DDoS:** dense-block detection (SVD, etc.), coreness
- **(F1, F2) Lockstep / collusion:** SVD, bridges, hubs, #incoming/outgoing calls

Cazzolato, M.T., Vijayakumar, S.; Lee, MC.; Vajiac, C.; Park, N.; Fidalgo, P.; Traina, A.J.M.; Faloutsos, C., *CallMine: Fraud Detectoin and Visualization of Million-Scale Call Graphs.* ACM CIKM, 2023.

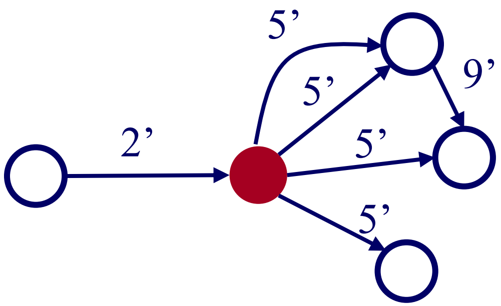
Specifically: Feature extraction

- How to turn nodes into n-dim vectors?
  - In-/out-degree (?)
  - In-/out calls (?)
  - In-/out minutes (?)
  - pageRank (?)
  - #triangles (?)
  - core-number (?)
  - Inter-arrival time (mean/median, IQR) (?)
  - Mean/median/IQR duration (?)



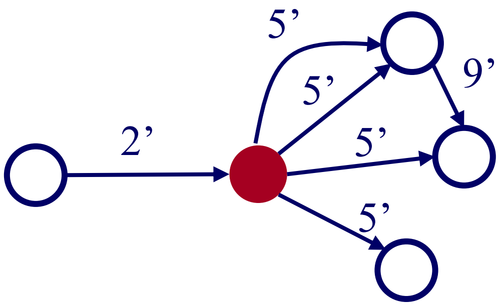
Specifically: Feature extraction

- How to turn nodes into n-dim vectors?
  - 👍 – In-/out-degree
  - 👍 – In-/out calls
  - 👍 – In-/out minutes
  - 👎 – pageRank
  - 👎 – #triangles
  - 👍 – core-number
  - 👍 – Inter-arrival time (mean/median, IQR)
  - 👍 – Mean/median/IQR duration



Specifically: Feature extraction

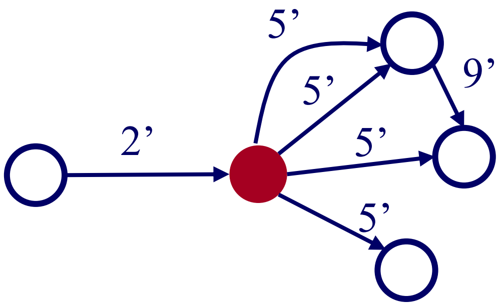
- **We cover:**
  - How <u>well connected</u> the node is
  - How many <u>distinct people</u> called
  - How many calls
  - Interval between calls (median, IQR)
  - Call duration (median, IQR)



Specifically: Feature extraction

- **We cover:**
  - How well connected the node is
  - How many distinct people called
  - How many calls
  - Interval between calls (median, IQR)
  - Call duration (median, IQR)

For incoming and outgoing calls

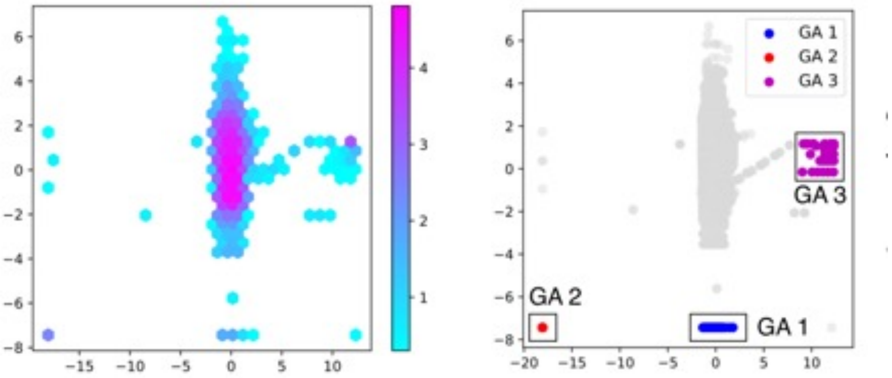


Bird's eye view

- …
- 4. Time evolving graphs
- 5. Visualization - practitioner's guide
  - Which features?
  - Outlier detection?
  - Visualization tools?
  - Case studies
- 6. Conclusions

Tools for outlier detection?

- LOF / MLOF
- Isolation Forests (scikit-learn)
- gen2Out: also spots micro-clusters
  - github



Lee, MC., Shekhar, S., Faloutsos, C., Hutson, TN., and Iasemidis, L., *gen2Out: Detecting and Ranking Generalized Anomalies.* IEEE Big Data, 2021.

Bird’s eye view

- …
- 4. Time evolving graphs
- 5. Visualization - practitioner’s guide
  - Which features?
  - Outlier detection?
  - Visualization tools?
  - Case studies
- 6. Conclusions

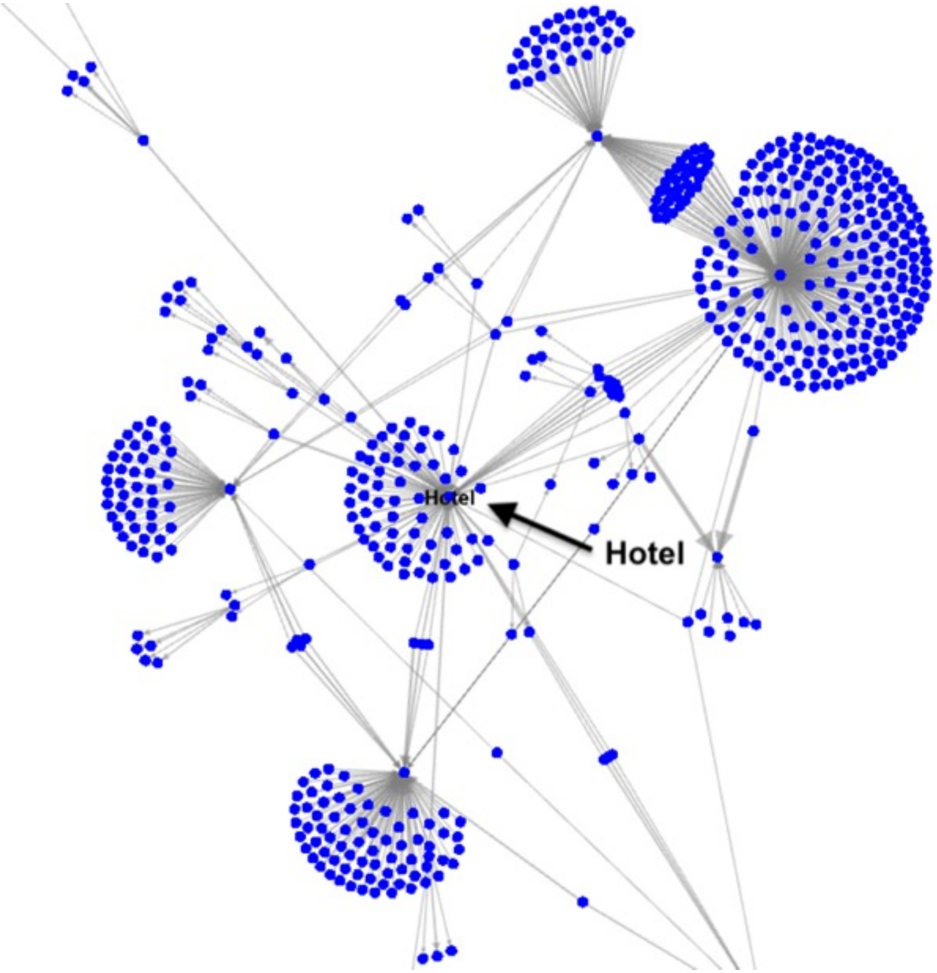
Visualization?

- Q: What to plot (for ~1M ~10-dim points)?

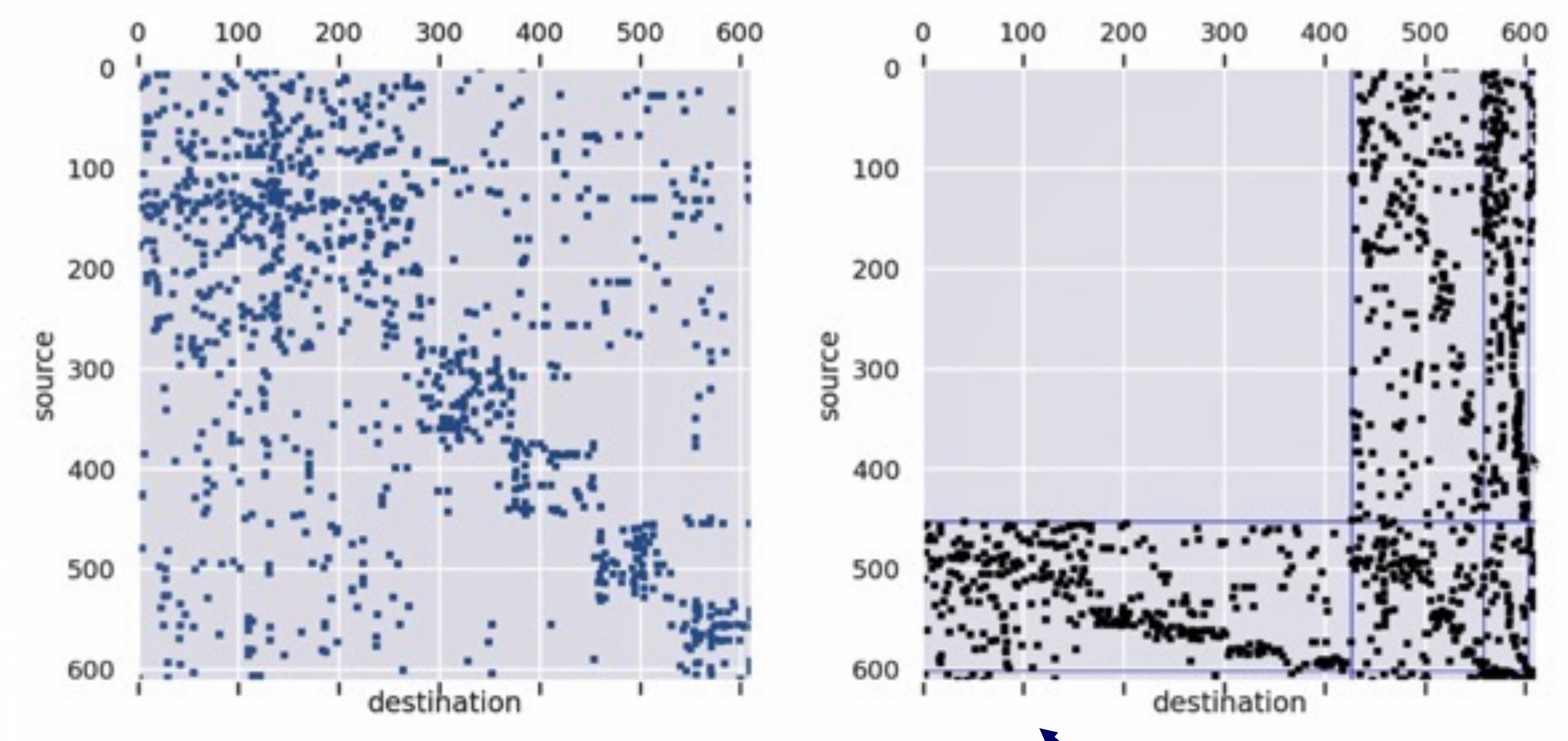
Visualization?

- Q: What to plot (for ~1M ~10-dim points)?
- **A1:** Spring model
- **A2:** Adjacency matrix
- **A3:** 1-d histograms (log-log)
- **A4:** 2-d scatter plots / heat maps (also log-log)
- **A5:** parallel coordinates
- **A6:** demo of TgraphSpot

A1: Spring model

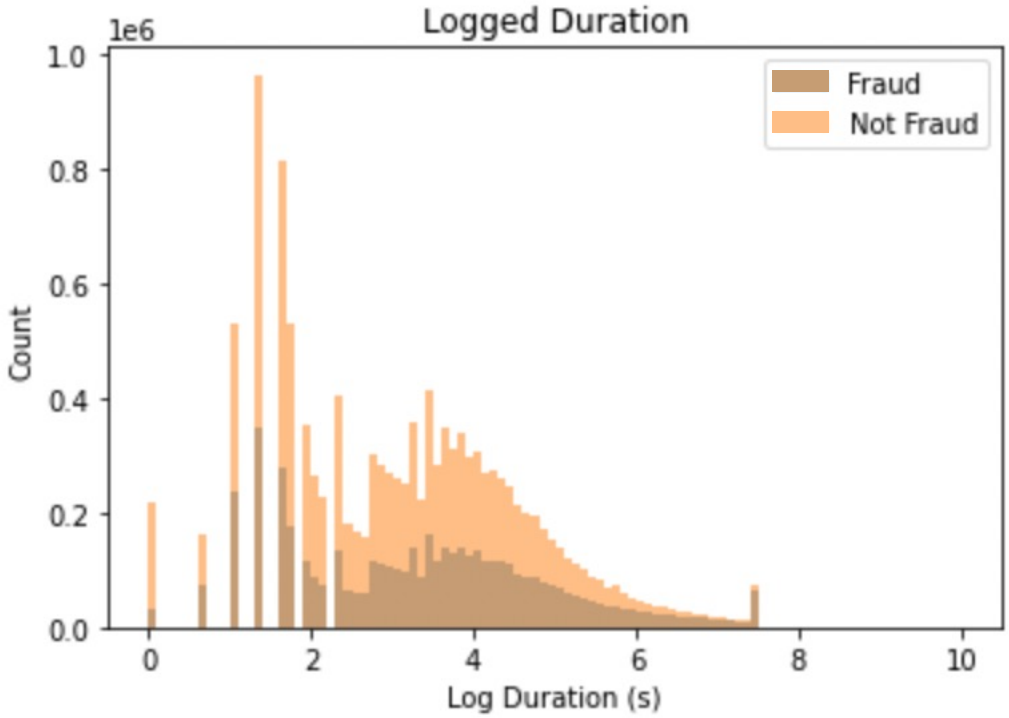


A2: Adjacency matrix (reordered)



Cross-associations

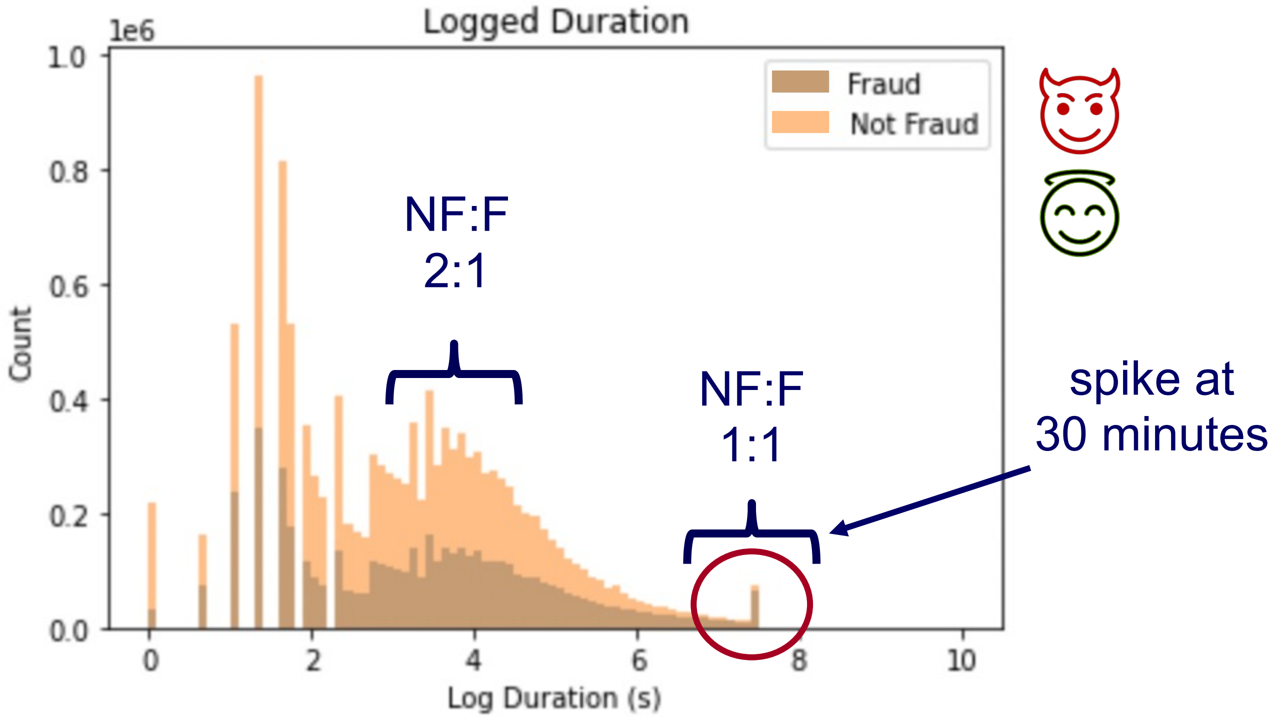
A3: 1d hist.: Phonecall durations



😈

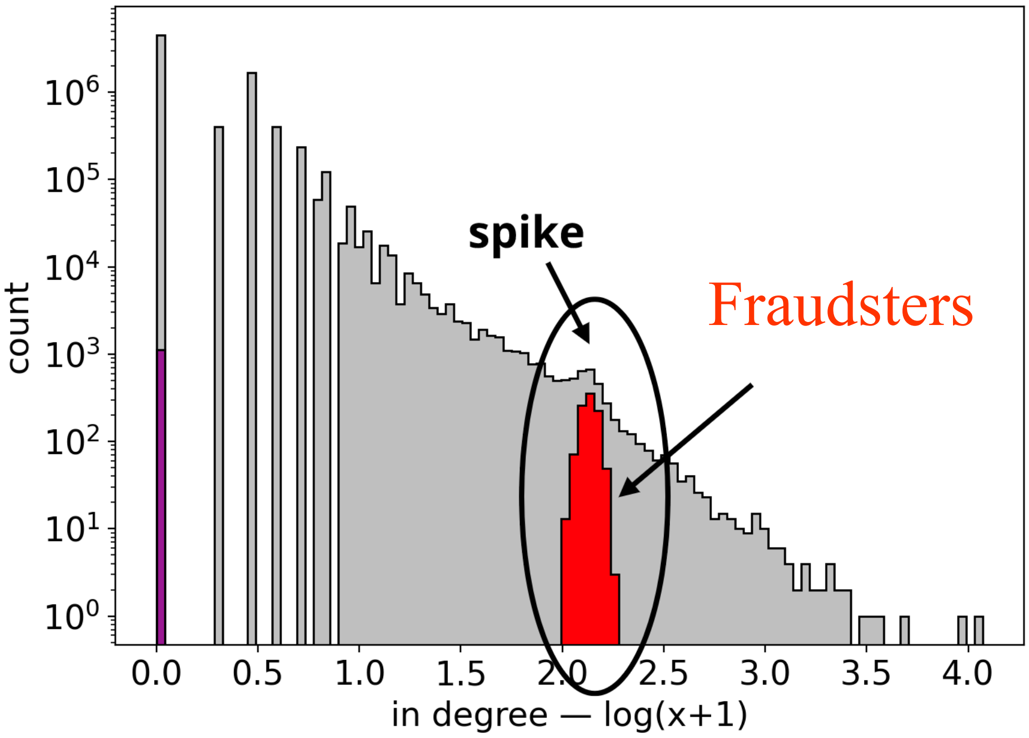
😇

A3: 1d hist.: Phonecall durations



- Duration distributions: comparable
- Spike @ 30’ (~50% fraudsters)

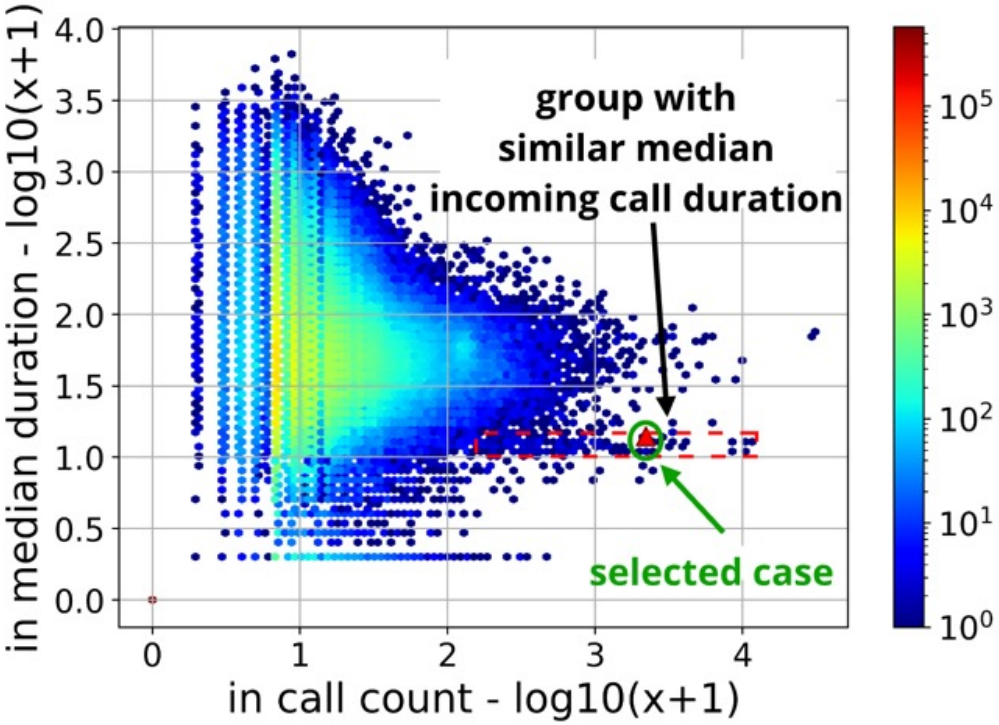
A3: 1d hist.: Phonecall durations



- Suspicious spike: #Incoming calls distribution

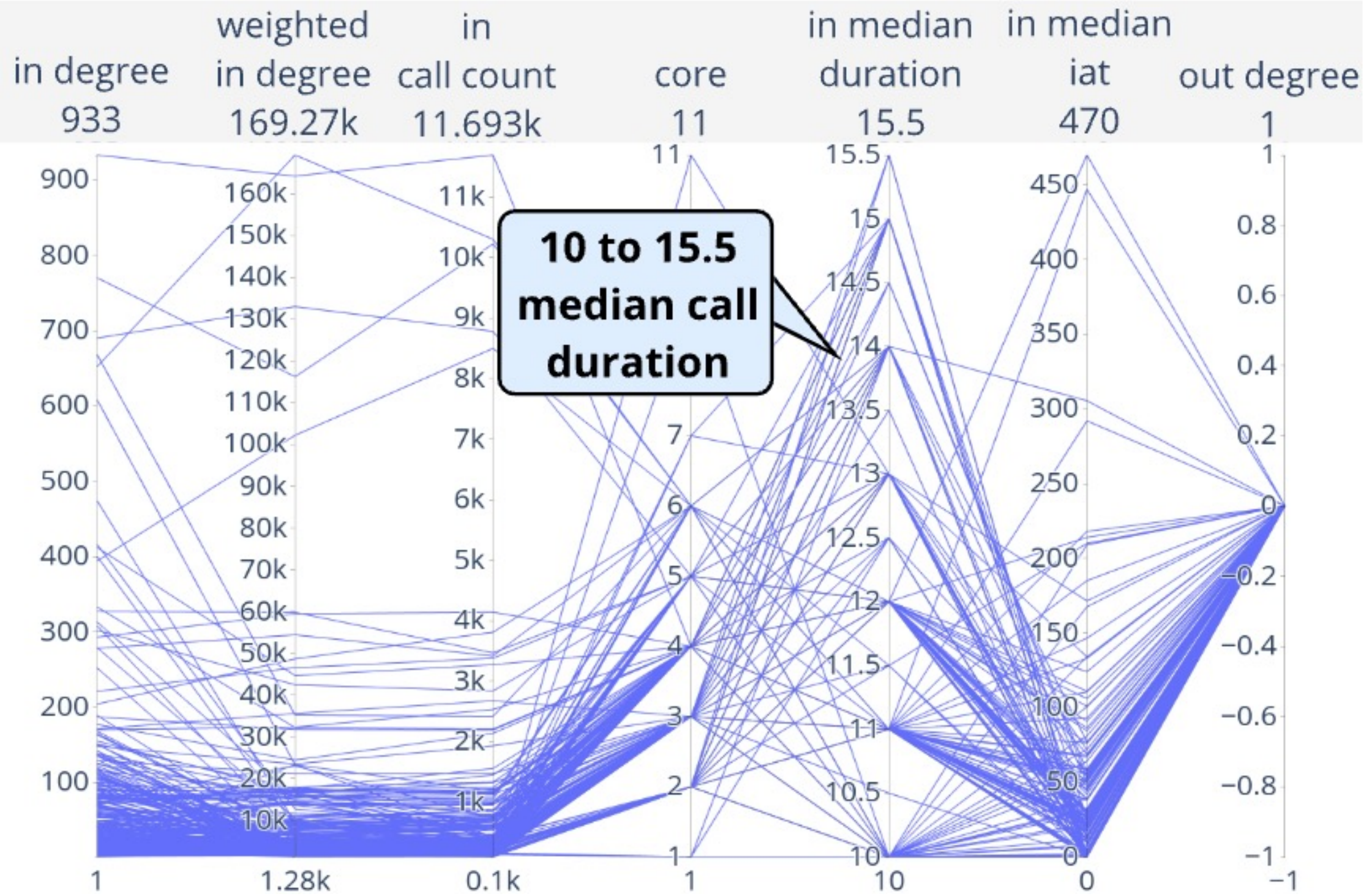
A4: 2-d heatmaps

Median in-duration



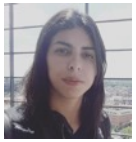
In-call count

A5: Parallel coordinates



10 to 15.5 median call duration

Recent tool: TgraphSpot



M. T. Cazzolato *et al.*, "TgraphSpot: Fast and Effective Anomaly Detection for Time-Evolving Graphs," *2022 IEEE Big Data*, 2022

Demo:

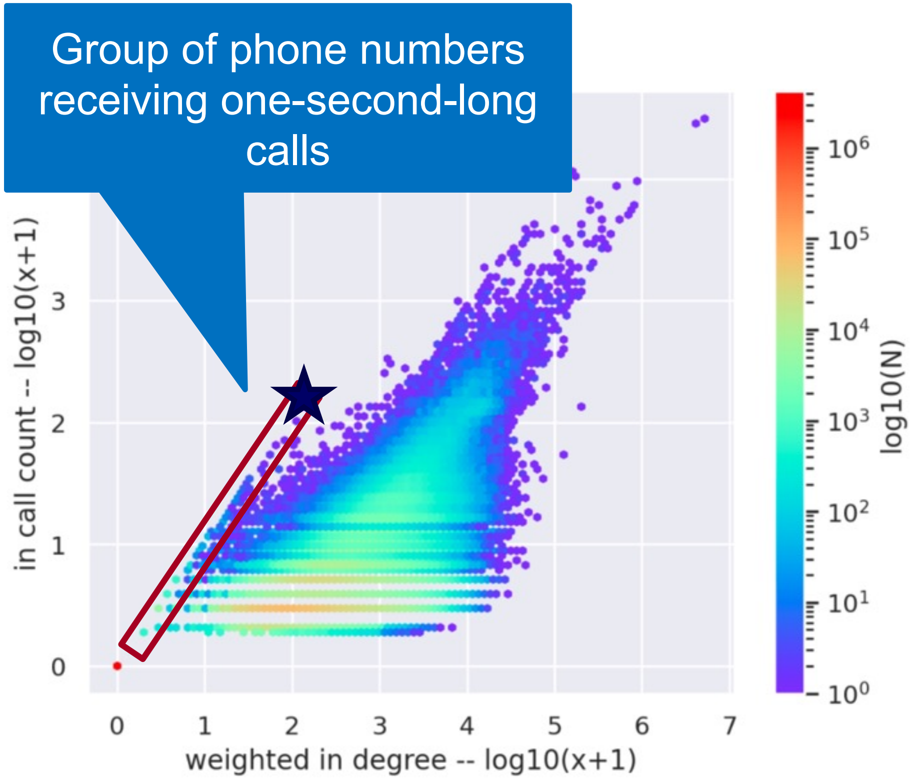
Github: https://github.com/mtcazzolato/tgraph-spot

(with video clips)

Bird’s eye view

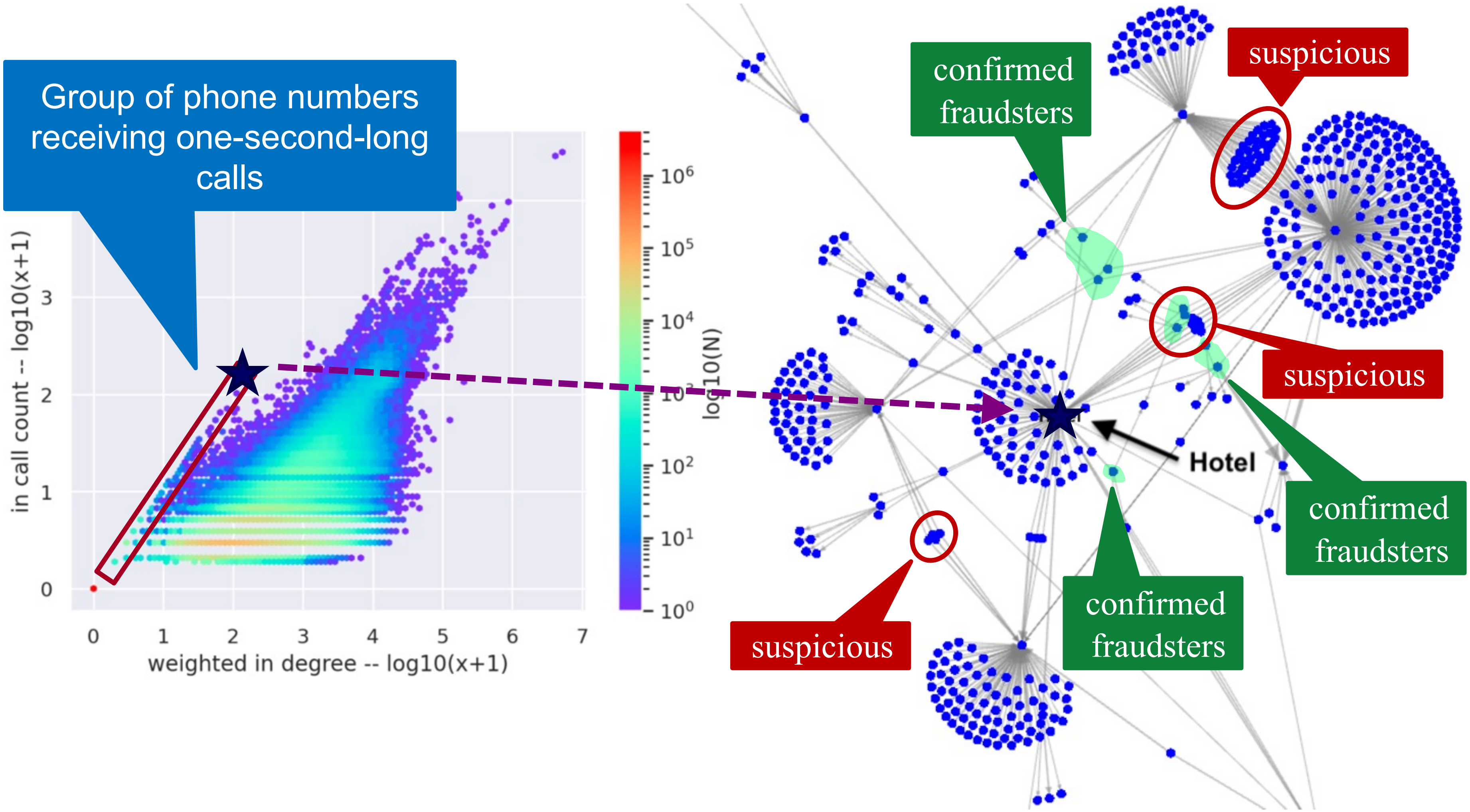
- …
- 4. Time evolving graphs
- 5. Visualization - practitioner’s guide
  - Which features?
  - Outlier detection?
  - Visualization tools?
  - Case studies
- 6. Conclusions

Case study #1



Group of phone numbers receiving one-second-long calls

Case study #1



Group of phone numbers receiving one-second-long calls

confirmed fraudsters

suspicious

suspicious

confirmed fraudsters

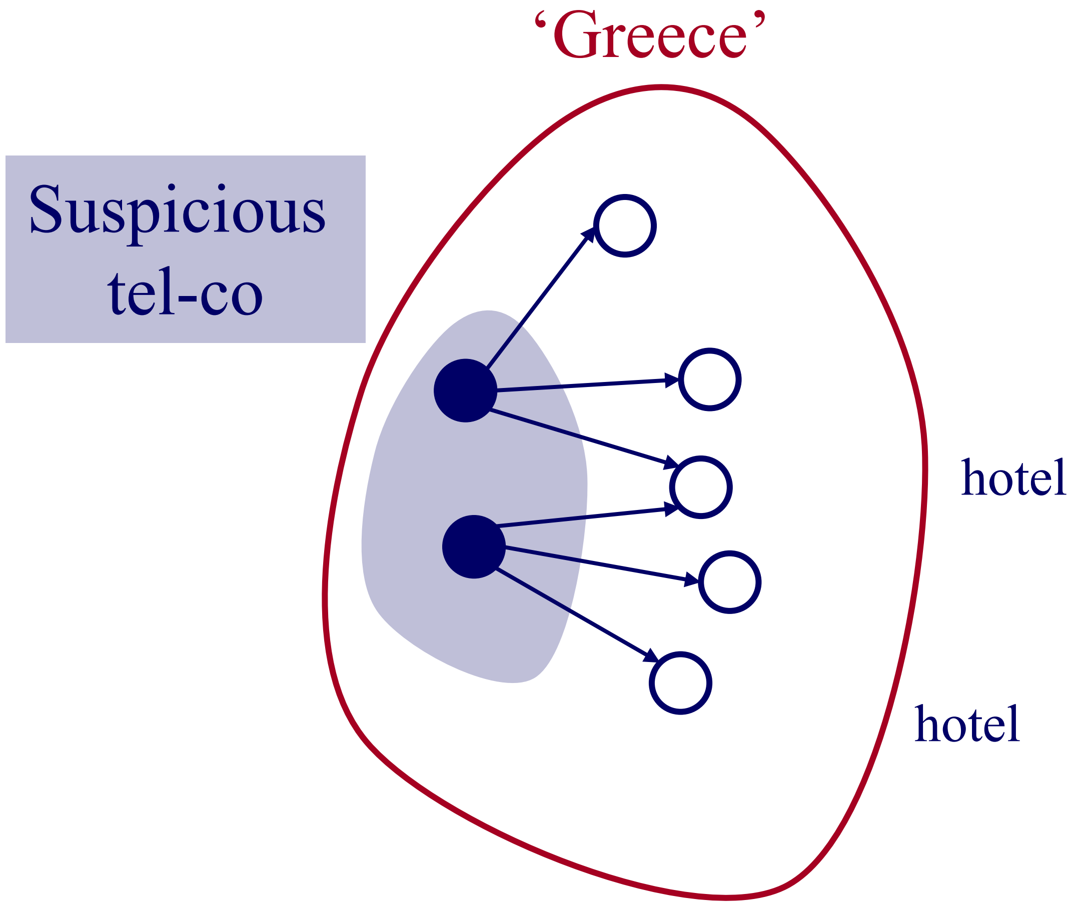
confirmed fraudsters

suspicious

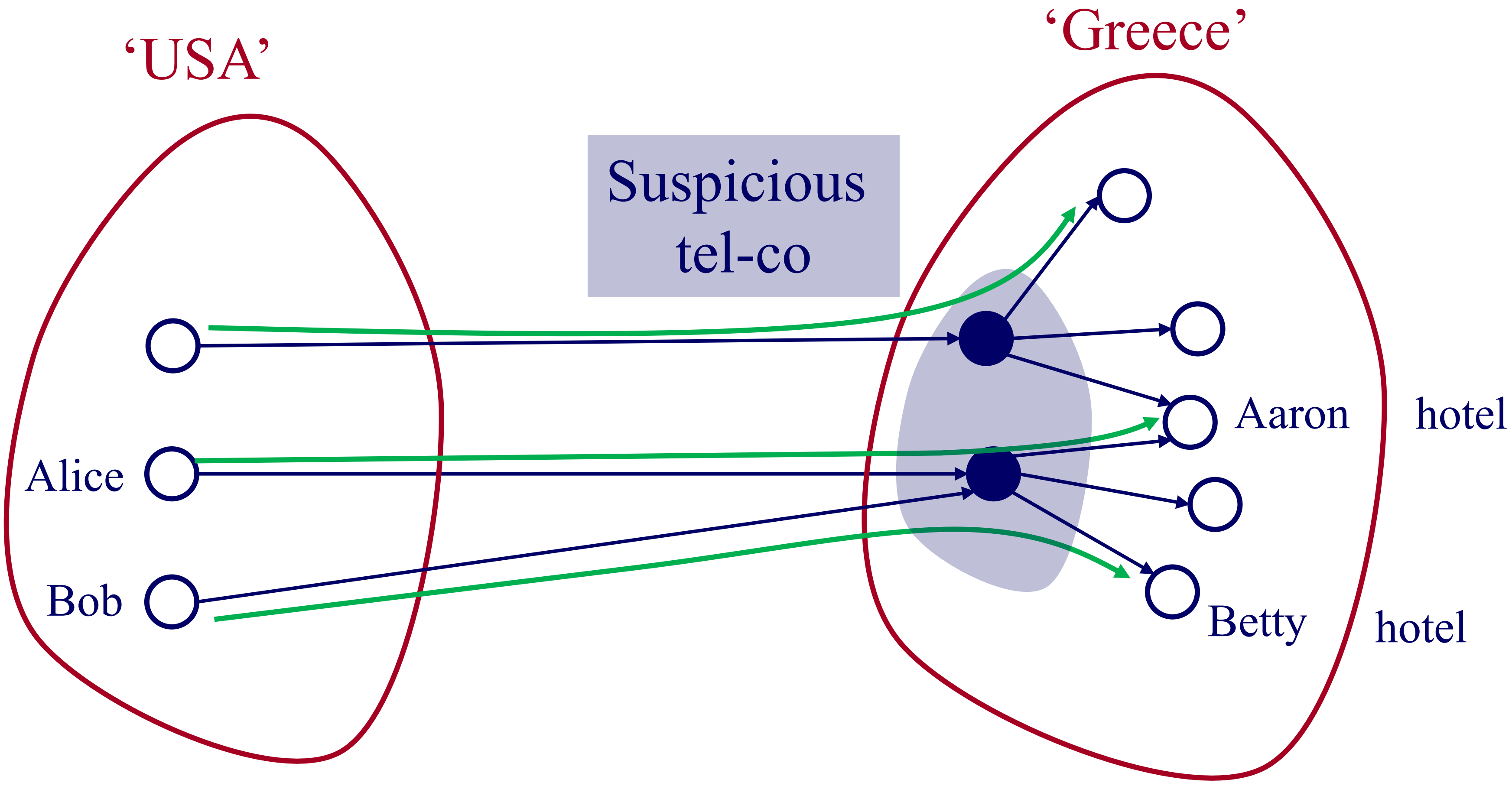
Q: Why?

- Q: Why would people call hotel-like numbers, for 1second?

A: ‘international by-pass’

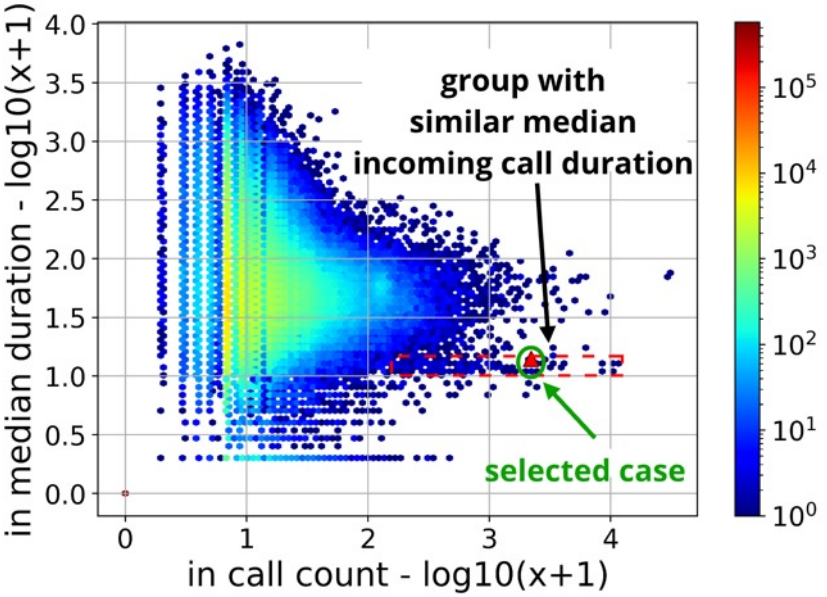


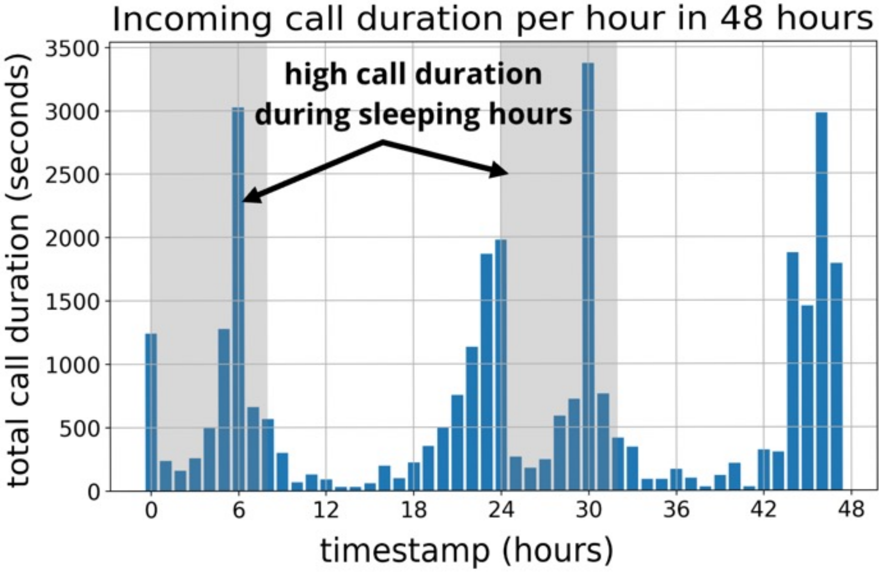
A: ‘international by-pass’



Case study #2

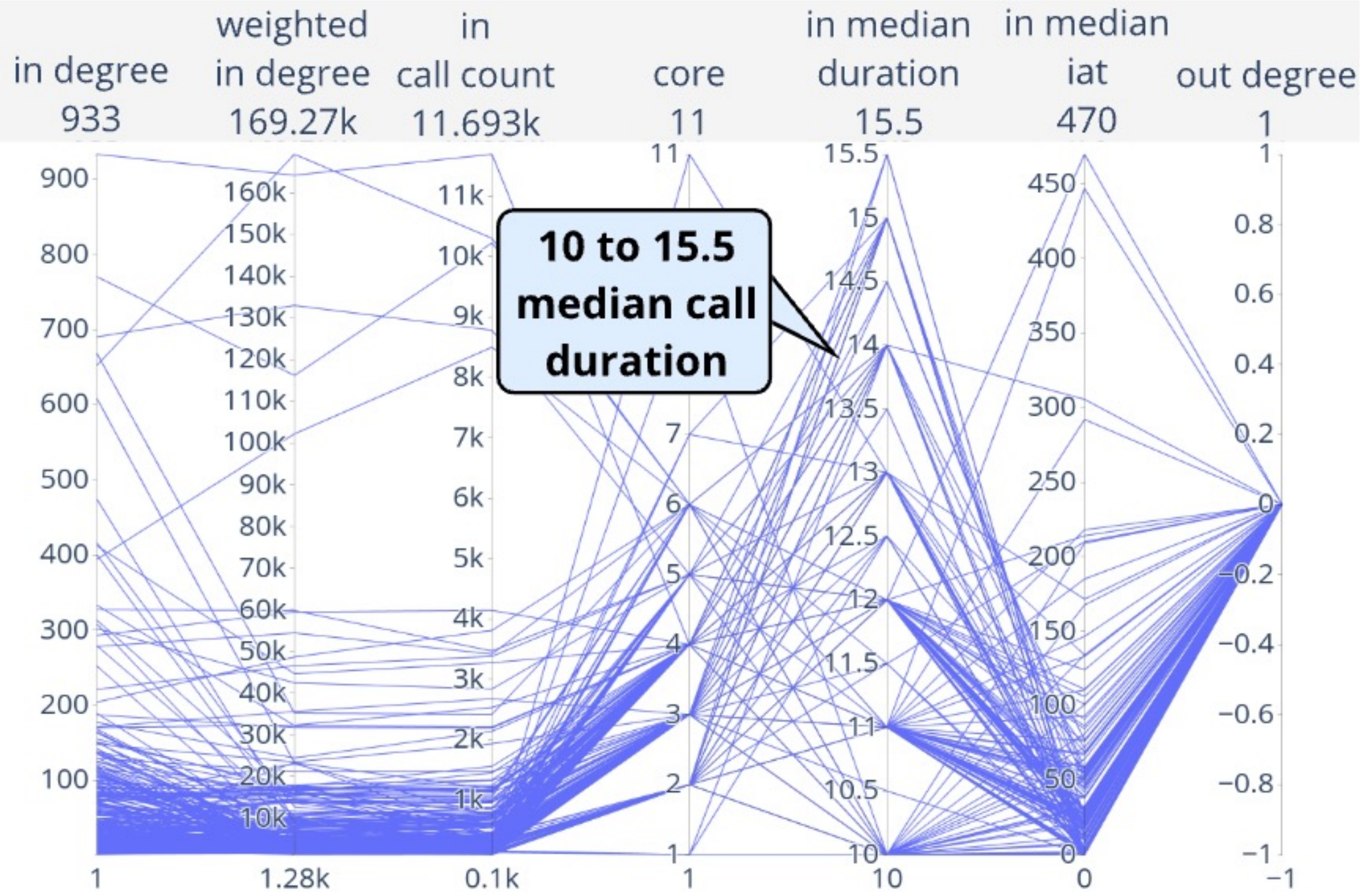
(median) duration





In-call count

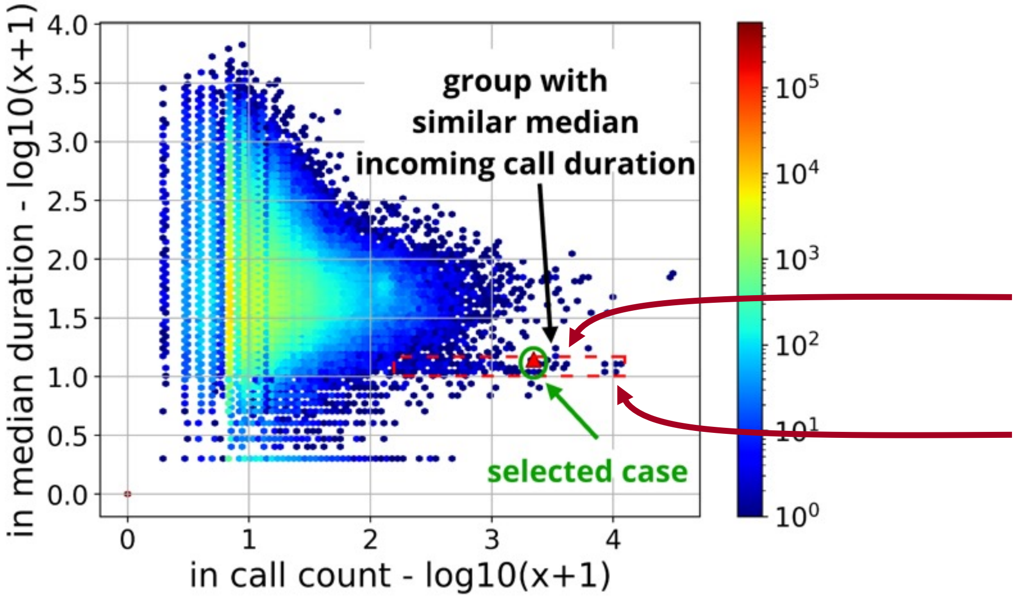
Case study #2



10 to 15.5 median call duration

Case study #2

(median) duration

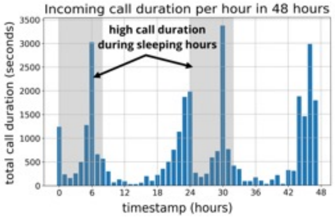


De-anonymization: Info numbers (weather, stocks, etc)

In call count

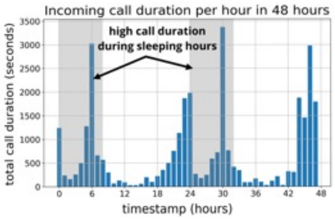
Case study #2 – ‘why?’

- Q: Why would someone call info numbers, 10’ at a time, during sleeping hours?

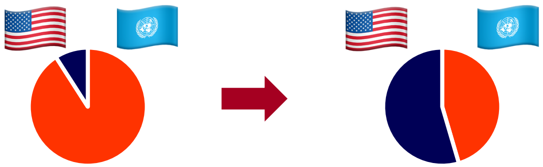


Case study #2 – ‘why?’

- Q: Why would someone call info numbers, 10’ at a time, during sleeping hours?



- A: ‘**camouflage**’:
  - The callers have a lot of (shady) international traffic
  - And call local numbers that won’t respond
  - So that the callers evade filters of ‘high fraction of international traffic’



‘Recipe’ Structure:

- Problem definition
- Short answer/solution
- LONG answer – details
- **Conclusion/short-answer**

Conclusions

Excellent tools for

- Static graphs (PR, SVD, BP, …)
- Time-evolving/het. graphs (tensors)

Visualization / explanations: vital

Christos Faloutsos

Pedro Fidalgo

Mirela Cazzolato