Machine learning approach to cross-device identification of users

Mateusz Jukiewicz[1] Bartek Bogacki[1] Krzysztof Dembczyński[2]

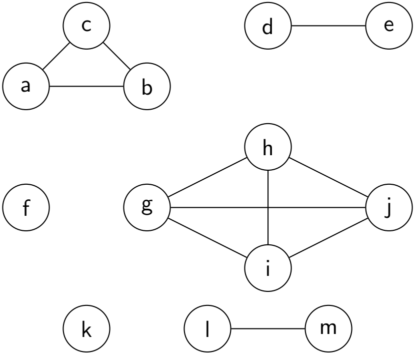
[1]Roq.ad GmbH, Germany
[2]Poznan University of Technology, Poland

IT Research Workshop at WCC 2018, Poznań, September 21, 2018

Cross-identification of users

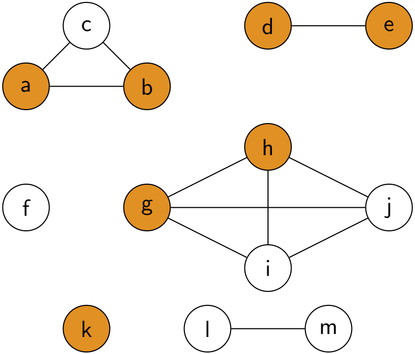
- **Applications**: Online advertising, content management and personalization, fraud detection.

Graph representation of users and devices



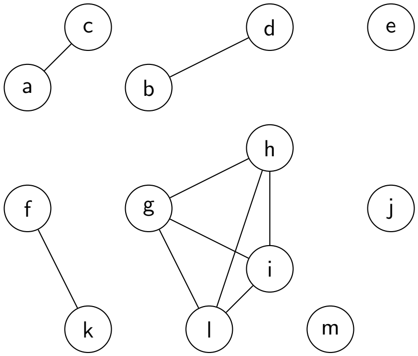
- nodes: devices (e.g., a, b, c, d, e, . . .)
- cliques: users with their devices (e.g., (a,b,c))

Deterministic cross-device graphs



- Unique factors to identify a person, e.g., email address or login name
- Quality far beyond from being perfect!
- Used for training and evaluating probabilistic solutions

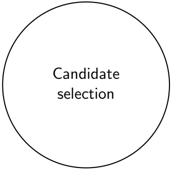
Probabilistic cross-device graphs



- Based on deep analysis of logs (behavior of devices in the Internet)
- Hand-made rule vs. Data-driven approach ($\Rightarrow$ Machine learning)

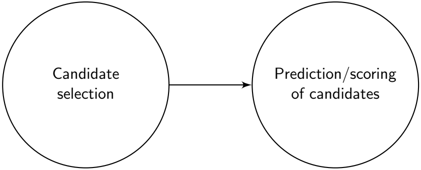
Standard machine learning approach for probabilistic graphs

Standard machine learning approach for probabilistic graphs



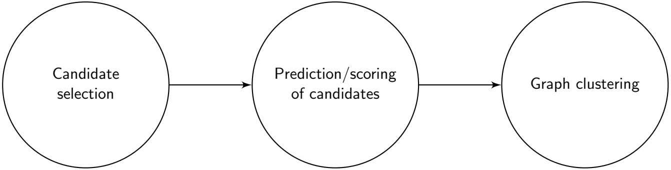
- Candidate selection: reducing the number of possible pairs by filtering them by some initial premises

Standard machine learning approach for probabilistic graphs



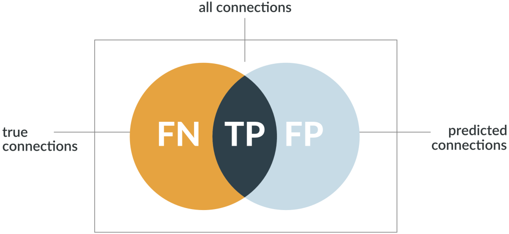
- Prediction/scoring: estimating the score for each candidate pair of devices

Standard machine learning approach for probabilistic graphs



- Graph clustering: construction of the probabilistic graph

Measuring performance of cross-device solutions



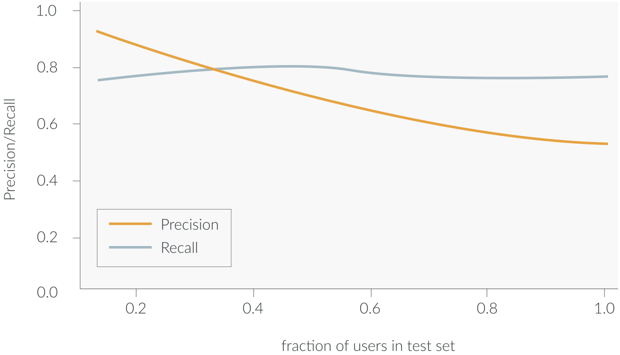
- Precision and recall:

$$\text{Recall} = P(\hat{y}=1|y=1) = \frac{P(y=1,\hat{y}=1)}{P(y=1)} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$
$$\text{Precision} = P(y=1|\hat{y}=1) = \frac{P(y=1,\hat{y}=1)}{P(\hat{y}=1)} = \frac{\text{TP}}{\text{TP}+\text{FP}},$$

where

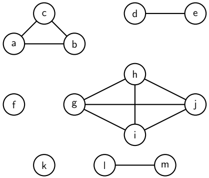
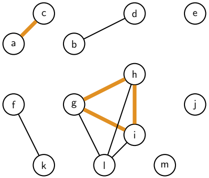
- $y = 1 \Rightarrow$ there exists a true connection between two devices,
- $\hat{y} = 1 \Rightarrow$ a connection has been predicted in the graph.

Pitfalls of the commonly used methodology

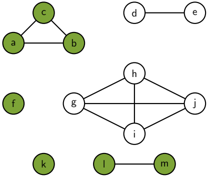
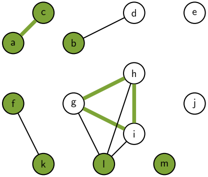
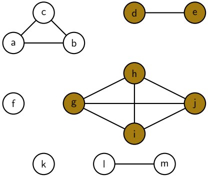
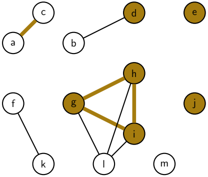


- Recall relatively stable with the size of deterministic graph
- Precision decreases with the size of deterministic graph (overestimation)!

Why precision decreases?

Ground truth graph	Probabilistic graph	Performance
		TP $= 4$ FP $= 5$ Precision $= 4/9$

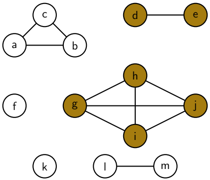
Why precision decreases?

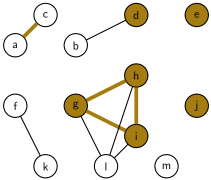
Deterministic graph	Induction of probabilistic graph	Performance
		$\mathrm{TP}_1 = 1$ $\mathrm{FP}_1 = 1$ $\mathrm{Precision}_1 = 0.5$
		$\mathrm{TP}_2 = 3$ $\mathrm{FP}_2 = 0$ $\mathrm{Precision}_2 = 1.0$

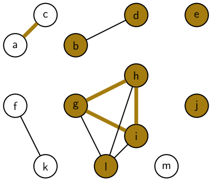
$\mathrm{TP} = 4 = \mathrm{TP}_1 + \mathrm{TP}_2 \quad \mathrm{FP} = 5 \neq \mathrm{FP}_1 + \mathrm{FP}_2 \quad \mathrm{Precision} = 4/9$

- Induction of probabilistic graph:

Deterministic graph	Lower induction of probabilistic graph	Upper induction of probabilistic graph



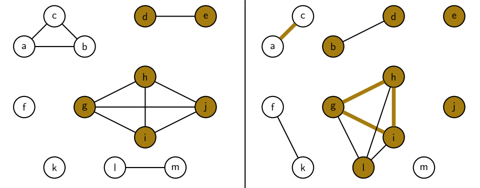




These two types of induction give the lower and upper bound of the value of precision.

- Device-based measures:
  - For each device $v \in V$ construct two lists:
    - $L(v)$: list of devices connected with v in deterministic graph,
    - $\hat{L}(v)$: list of devices connected with v in probabilistic graph.

Deterministic graph	Upper induction of probabilistic graph



$$L(g) = [h, i, j] \quad \hat{L}(g) = [h, i, l]$$

- Device-based measures:
  - The performance is then averaged over single devices:

$$M_V = \frac{1}{|V|} \sum_{v \in V} M_v(L(v), \hat{L}(v))\,.$$

Towards new standards

- Device-based measures:
  - The performance is then averaged over single devices:

$$M_V = \frac{1}{|V|} \sum_{v \in V} M_v(L(v), \hat{L}(v)) \,.$$

  - M_v can be defined, for example, as a device-based recall and precision:

$$\begin{aligned} \mathrm{Recl}(L(v), \hat{L}(v)) &= \frac{|L(v) \cap \hat{L}(v)|}{|L(v)|} \,, \\ \mathrm{Prec}(L(v), \hat{L}(v)) &= \frac{|L(v) \cap \hat{L}(v)|}{|\hat{L}(v)|} \,. \end{aligned}$$

Summary

- Cross-device identification — actual and challenging problem.
- Machine learning approach to cross-device identification.
- Measuring performance of cross-device identification solutions.

Summary

- Cross-device identification — actual and challenging problem.
- Machine learning approach to cross-device identification.
- Measuring performance of cross-device identification solutions.

Thank you

Q&A