MUNI
C4E

Cyber Key Terrain Identification Using Adjusted PageRank Centrality

Lukáš Sadlek, Pavel Čeleda
sadlek@mail.muni.cz

Masaryk University, Czech Republic

June 16, 2023 @ IFIP SEC 2023 conference

Introduction

Motivation

- Cyber defense is meaningless **without** knowing **which cyber assets** to protect
- Protect cyber assets **related** to the organization's **objectives** (mission)
- Verify the **content** of populated **asset inventory**

Cyber Key Terrain

- Network **devices**, network **services**, cyber **personas**, and other **network entities**
- Provides an **advantage** for attackers and defenders
- **Example** key asset can be a local **domain name server**

Research Questions

RQ1 *How to determine **which IP addresses** from cyber terrain are the **key** according to the **network communication**?*

Research Questions

RQ1 *How to determine **which IP addresses** from cyber terrain are the **key** according to the **network communication**?*

RQ2 *Does **adjusting** the PageRank centrality lead to **better correctness** of determining the cyber key terrain, and can it process IP flows from the **real-world network**?*

Network Centrality

Network Centrality Measures

- Asset **criticality** based on **position** in a graph
- **Several types** – degree, betweenness, closeness, and eigenvector centralities
- **Data** – IP flows (unidirectional and bidirectional)

PageRank Centrality

- Initially proposed for **ranking of web pages**
- Considers vertices linked by **outgoing edges**
- **Damping factor** – the probability of continuing with a **random web page**

PageRank Centrality – Example

Important Nodes

- Linked by **many** nodes
- Linked by an **important node** that references a **small** number of other nodes

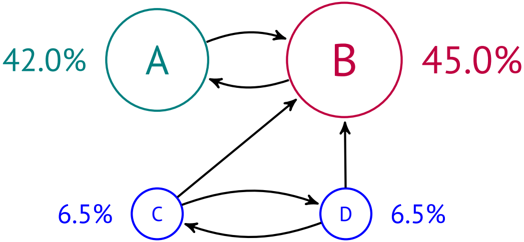

Figure 1: **PageRank centrality** expressed **in percentages** for an example graph.

Learning Phase – I

Adjusted PageRank Formula

- Asset criticality by considering **network communication specifics**
- Damping factors **adjusted** to source and destination **port pairs**
- **All** adjusted PageRank values **sum to one**

Machine Learning Methods

- **Hill climbing** – modification of the **current** solution to achieve a **better** solution
- **Random walk** – assignment of a **random** value to a **conflict** variable
- **Optimization** problems

Learning Phase – II

Algorithm 1: Learning phase

Input : graph, max_iterations, probability, results, heuristic
Output: best F1 score, best damping factors

```
preprocessing()
one_iteration_of_pagerank(graph, factors, results)
iterations ← 0
while F1_score ≠ 1 and iterations ≤ max_iterations do
    assign_best_F1_score()
    port_pair ← choose_random_conflict_port_pair()
    if random_experiment > 1 − probability then
        factors[port_pair] ← random(0, 1)
    end
    else hill_climbing(heuristic)
    iterations ← iterations + 1
    one_iteration_of_pagerank(graph, factors, results)
end
assign_best_F1_score()
```

Computation Phase

Dynamic Stream-Based PageRank

- **Multiple** damping factors
- **One** IP flow (**forward** direction of **bidirectional** flow) = **one** edge
- **Values** of PageRank centrality **fluctuate** throughout the time

Advantages

- Processes a **large** amount of data
- Reads **sorted** flows in **one pass**

Evaluation – Dataset from Cyber Defense Exercise - I

Methodology

- Six participating teams, **six** partial **datasets**

		Team 1	Team 2	Team 3	Team 4	Team 5	Team 6
Data	Nodes	554	1,380	542	884	503	219
	Edges	1,468	3,064	1,631	2,418	1,361	584
	IP flows	66,499	116,897	63,400	88,734	78,254	30,781
Time	Preprocessing	0.5 s	1.4 s	0.4 s	0.9 s	0.4 s	0.1 s
	Learning time	175.0 s	**17.5 min**	168.3 s	7.4 min	142.9 s	30.2 s
	Computation time	1.3 s	2.1 s	1.2 s	1.8 s	1.6 s	0.6 s

Table 1: The **size** of the processed graph for learning and measured **time**.

Evaluation – Dataset from the Cyber Defense Exercise – II

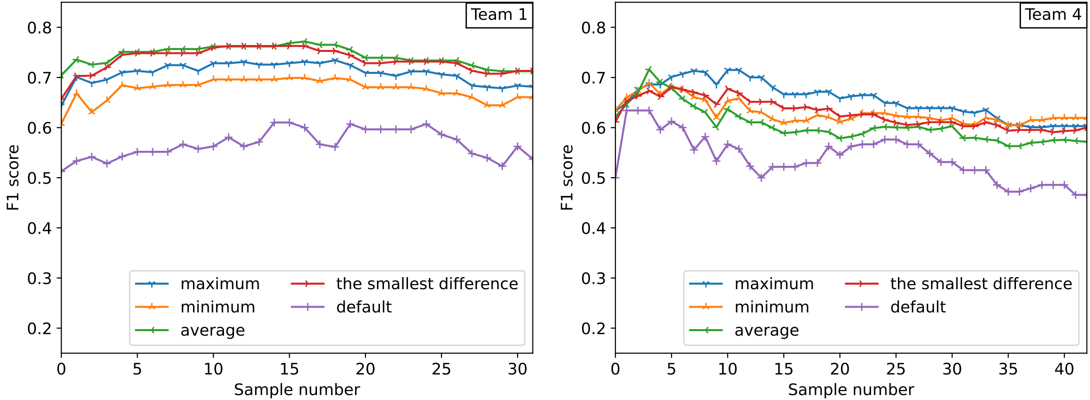

Figure 2: Line graphs containing **F1 scores for heuristics** and the PageRank with **default damping factor** for two team networks.

Evaluation – Dataset from the Campus Network – I

Heuristic	S1	S2	S3	S4	S5	N	Var
Default PageRank	24	24	23	23	23	30.8	–
Minimum	**26.4**	**28.3**	**27.6**	**27.9**	**27.8**	**37.9**	**18.9**
Maximum	25.2	25.8	24.5	24.2	24.2	34.1	13.6
Average	22.8	23.3	22.3	22.6	22.4	29.7	10.3
The smallest difference	24.0	24.5	23.5	23.4	23.4	32.8	7.72

Table 2: The number of true positives in the **top 100 results** according to samples (S1 – S5), the average number of hosts from the **university network** (N), and the average variance of true positives during the **ten-minute-long window**.

Evaluation – Dataset from the Campus Network – II

Heuristic	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**N**	**Var**
Default PageRank	**41**	41	**43**	**42**	**41**	**41**	**41**	**41**	59.1	–
Minimum	40.3	41.5	41.1	40.7	40.1	40.1	40.3	39.9	**59.2**	2.7
Maximum	40.5	**41.7**	41.2	40.8	40.4	39.9	39.9	39.3	58.3	**11.2**
Average	38.6	39.8	39.7	39.1	38.7	38.0	38.6	38.3	56.9	7.8
The smallest difference	40.4	41.4	41.1	40.6	40.5	40.3	40.2	39.9	58.4	3.8

Table 3: The number of true positives in the **top 100 results** according to samples (S1 – S8), the average number of hosts from the **university network** (N), and the average variance of true positives in these samples during the **one-hour-long window**.

Limitations

Method

- Consider **other attributes** of IP flows
- IP flows may **not be optimally** sorted
- Results may **not fit** into the main memory

Evaluation

- **Progress** could remain **hidden**

Summary

Contribution

- **PageRank centrality** – IP addresses related to **critical organization's services**
- The **top 100 results** – almost the **same** precision with the **increased count** of flows
- **Approach** – easy **update** of values, a small number of **manual steps**

Supplementary Materials

- A proof-of-concept **implementation** of the learning and computation **phases**
- **Ground-truth** labels
- Available at https://doi.org/10.5281/zenodo.7884228

MUNI
C4E

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

C4E.CZ