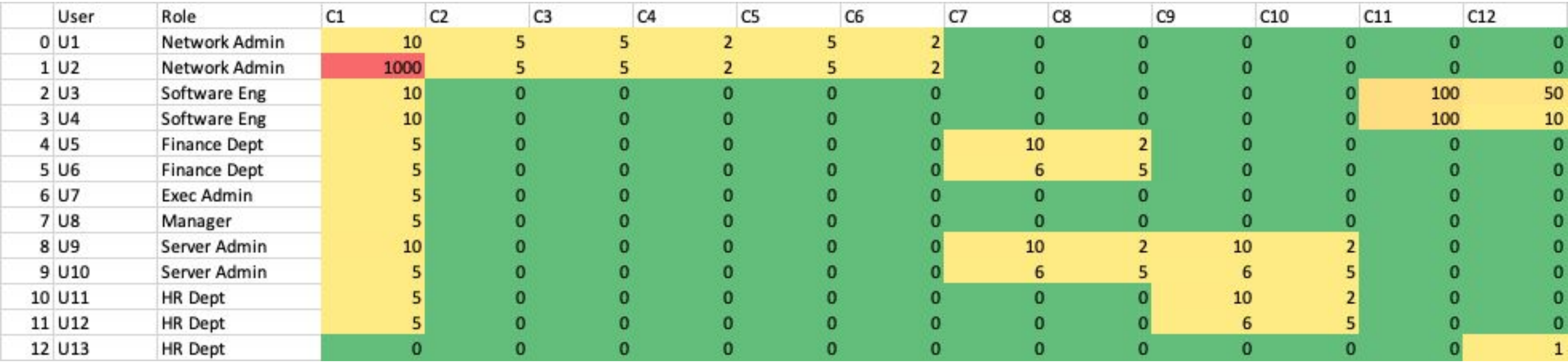
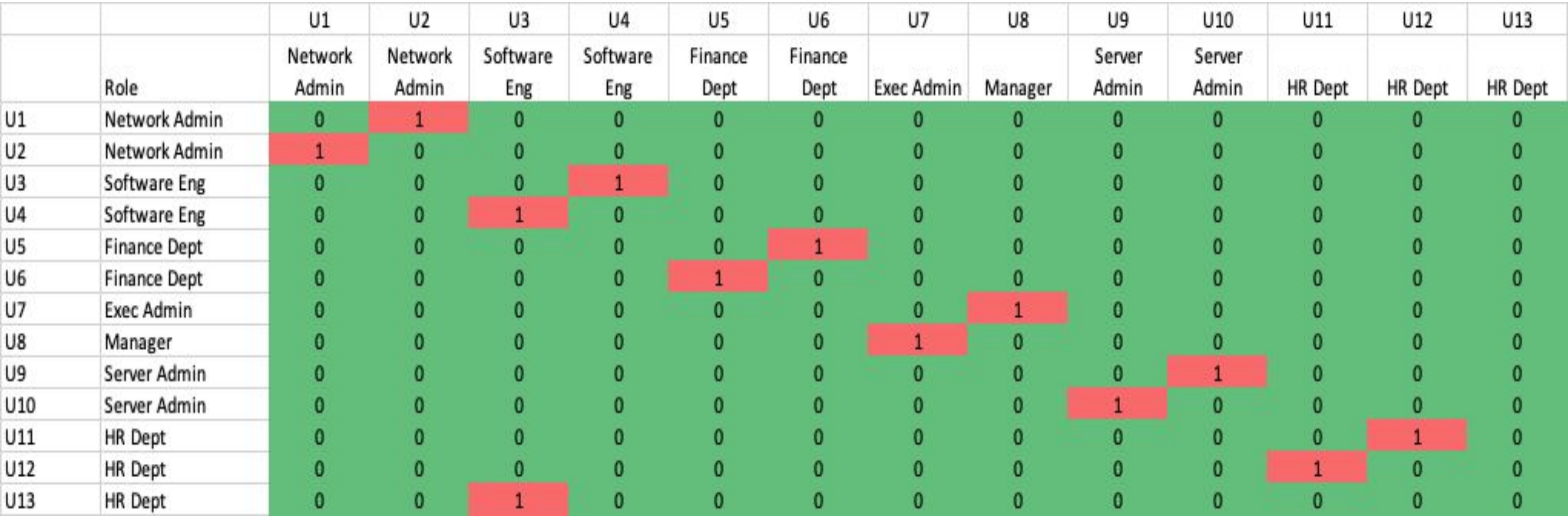
Peer Grouping

Initial dataset



	User	Role	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
0	U1	Network Admin	10	5	5	2	5	2	0	0	0	0	0	0
1	U2	Network Admin	1000	5	5	2	5	2	0	0	0	0	0	0
2	U3	Software Eng	10	0	0	0	0	0	0	0	0	0	100	50
3	U4	Software Eng	10	0	0	0	0	0	0	0	0	0	100	10
4	U5	Finance Dept	5	0	0	0	0	0	10	2	0	0	0	0
5	U6	Finance Dept	5	0	0	0	0	0	6	5	0	0	0	0
6	U7	Exec Admin	5	0	0	0	0	0	0	0	0	0	0	0
7	U8	Manager	5	0	0	0	0	0	0	0	0	0	0	0
8	U9	Server Admin	10	0	0	0	0	0	10	2	10	2	0	0
9	U10	Server Admin	5	0	0	0	0	0	6	5	6	5	0	0
10	U11	HR Dept	5	0	0	0	0	0	0	0	10	2	0	0
11	U12	HR Dept	5	0	0	0	0	0	0	0	6	5	0	0
12	U13	HR Dept	0	0	0	0	0	0	0	0	0	0	0	1

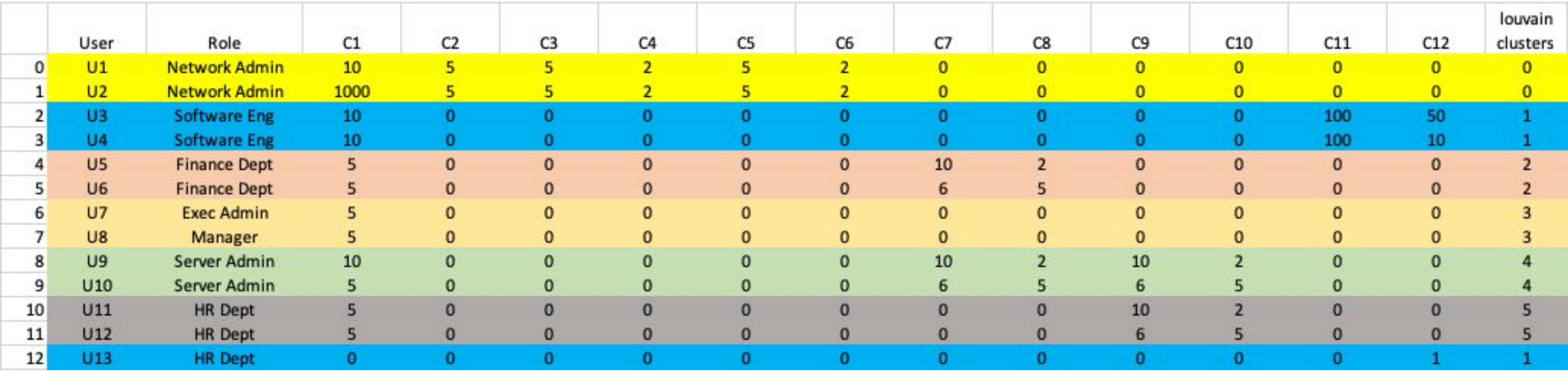
Adjacency matrix



		U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13
	Role	Network Admin	Network Admin	Software Eng	Software Eng	Finance Dept	Finance Dept	Exec Admin	Manager	Server Admin	Server Admin	HR Dept	HR Dept	HR Dept
U1	Network Admin	0	1	0	0	0	0	0	0	0	0	0	0	0
U2	Network Admin	1	0	0	0	0	0	0	0	0	0	0	0	0
U3	Software Eng	0	0	0	1	0	0	0	0	0	0	0	0	0
U4	Software Eng	0	0	1	0	0	0	0	0	0	0	0	0	0
U5	Finance Dept	0	0	0	0	0	1	0	0	0	0	0	0	0
U6	Finance Dept	0	0	0	0	1	0	0	0	0	0	0	0	0
U7	Exec Admin	0	0	0	0	0	0	0	1	0	0	0	0	0
U8	Manager	0	0	0	0	0	0	1	0	0	0	0	0	0
U9	Server Admin	0	0	0	0	0	0	0	0	0	1	0	0	0
U10	Server Admin	0	0	0	0	0	0	0	0	1	0	0	0	0
U11	HR Dept	0	0	0	0	0	0	0	0	0	0	0	1	0
U12	HR Dept	0	0	0	0	0	0	0	0	0	0	1	0	0
U13	HR Dept	0	0	1	0	0	0	0	0	0	0	0	0	0

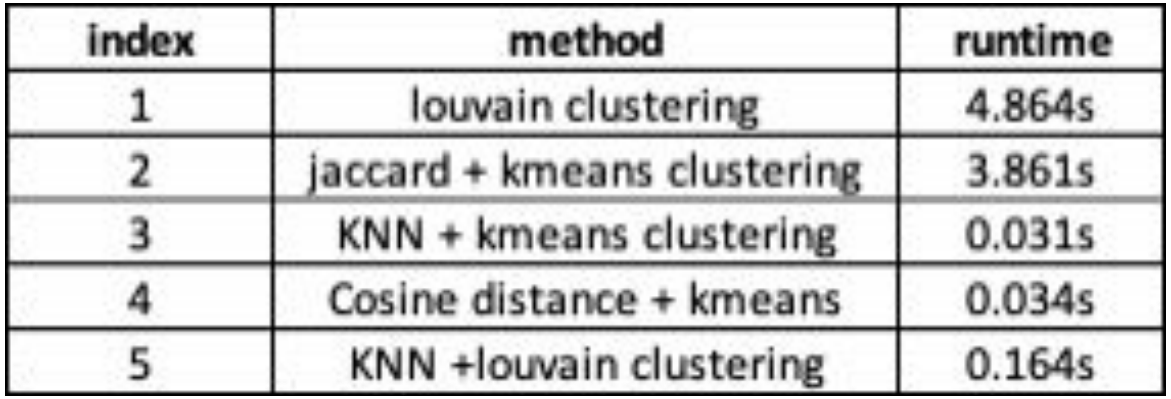
Neighborhood is computed using Euclidian distance but other distance measures can also be considered

Louvain clustering



	User	Role	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	louvain clusters
0	U1	Network Admin	10	5	5	2	5	2	0	0	0	0	0	0	0
1	U2	Network Admin	1000	5	5	2	5	2	0	0	0	0	0	0	0
2	U3	Software Eng	10	0	0	0	0	0	0	0	0	0	100	50	1
3	U4	Software Eng	10	0	0	0	0	0	0	0	0	0	100	10	1
4	U5	Finance Dept	5	0	0	0	0	0	10	2	0	0	0	0	2
5	U6	Finance Dept	5	0	0	0	0	0	6	5	0	0	0	0	2
6	U7	Exec Admin	5	0	0	0	0	0	0	0	0	0	0	0	3
7	U8	Manager	5	0	0	0	0	0	0	0	0	0	0	0	3
8	U9	Server Admin	10	0	0	0	0	0	10	2	10	2	0	0	4
9	U10	Server Admin	5	0	0	0	0	0	6	5	6	5	0	0	4
10	U11	HR Dept	5	0	0	0	0	0	0	0	10	2	0	0	5
11	U12	HR Dept	5	0	0	0	0	0	0	0	6	5	0	0	5
12	U13	HR Dept	0	0	0	0	0	0	0	0	0	0	0	1	1

Comparing runtimes of various methods



index	method	runtime
1	louvain clustering	4.864s
2	jaccard + kmeans clustering	3.861s
3	KNN + kmeans clustering	0.031s
4	Cosine distance + kmeans	0.034s
5	KNN +louvain clustering	0.164s

Runtimes based on 60 users and 14 computers

- Louvain outperforms other clustering techniques on accuracy and repeatability
- Louvain on its own is the slowest in runtime
- Combining Louvain with KNN improves runtime significantly without giving up much in accuracy

Louvain algorithm procedure

1) collect user-computer access logs pertaining to successful logins [User_ID,Computer_ID/IP]

2) Drop repeated events

3) Use KNN to identify closest neighbor(s) to each user

3) Use nearest neighbors to form an adjacency matrix a_m

4) Use a_m as input for Louvain clustering

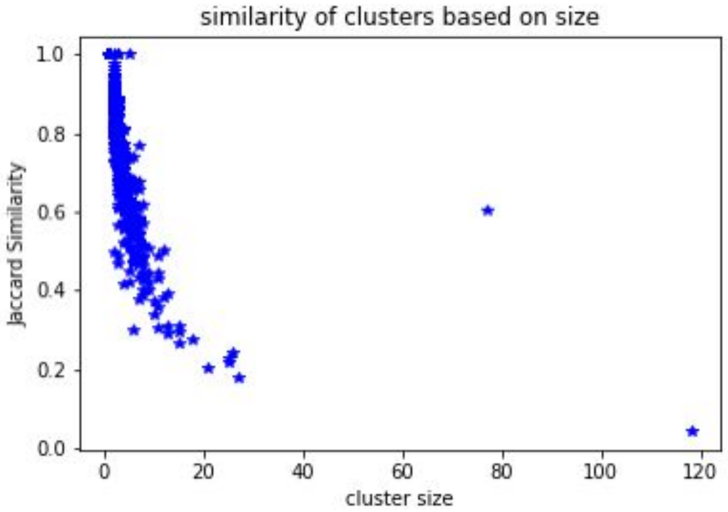
5) Obtain clusters from Louvain routine

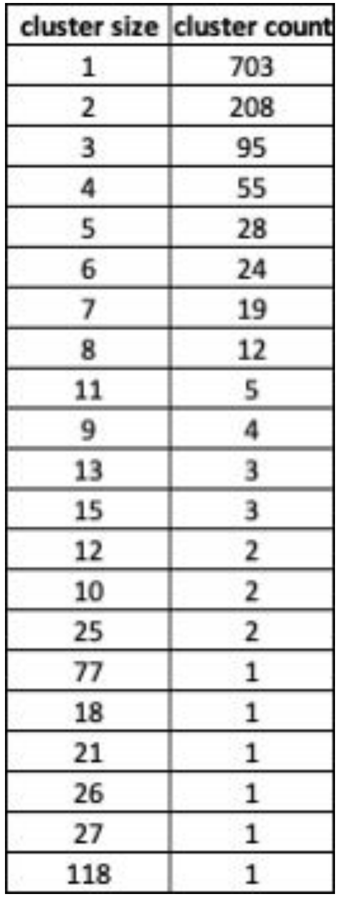
6) If there are left over users un-clustered go back to step(3) and repeat until all users are clustered

Los Alamos dataset

- 1 day of los Alamos data = 2.3 M authentication events, 33K users and 9.5K computers
- Filtered out events – users that have accessed < 5 machines, computers accessed by < 3 users – how many left ??
- We used Jaccard distance to compute tightness of clusters
- It took 7 iterations using Louvain after which about 4 users were left un-clustered.
- Starting with higher number of neighbors – leads to bigger clusters and less iterations to run Louvain.

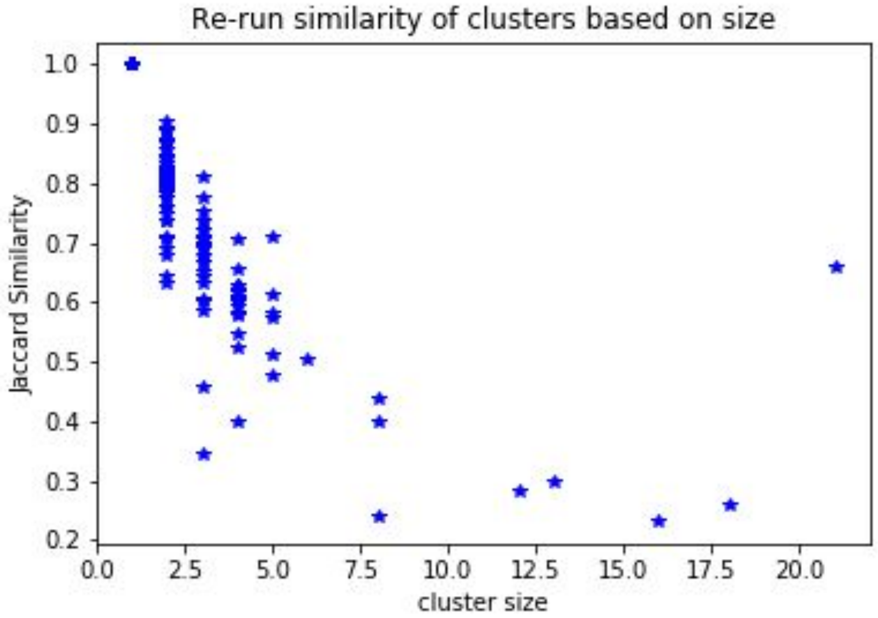
Round 1 clustering results

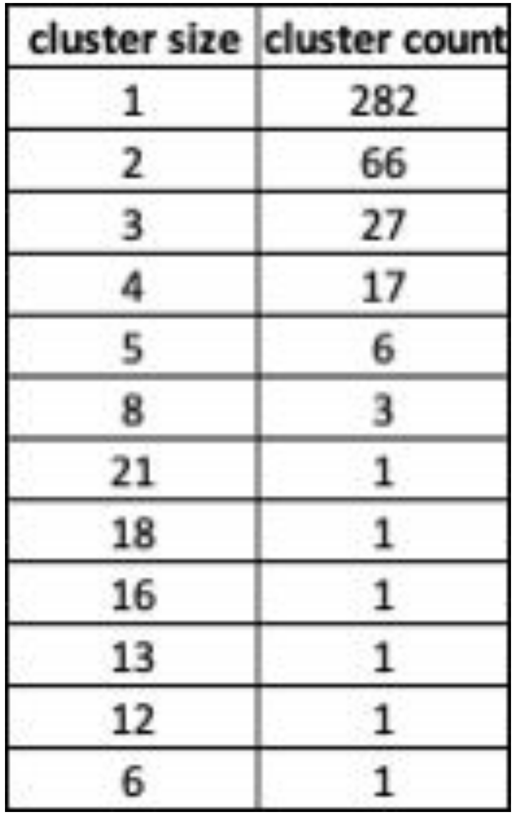




cluster size	cluster count
1	703
2	208
3	95
4	55
5	28
6	24
7	19
8	12
11	5
9	4
13	3
15	3
12	2
10	2
25	2
77	1
18	1
21	1
26	1
27	1
118	1

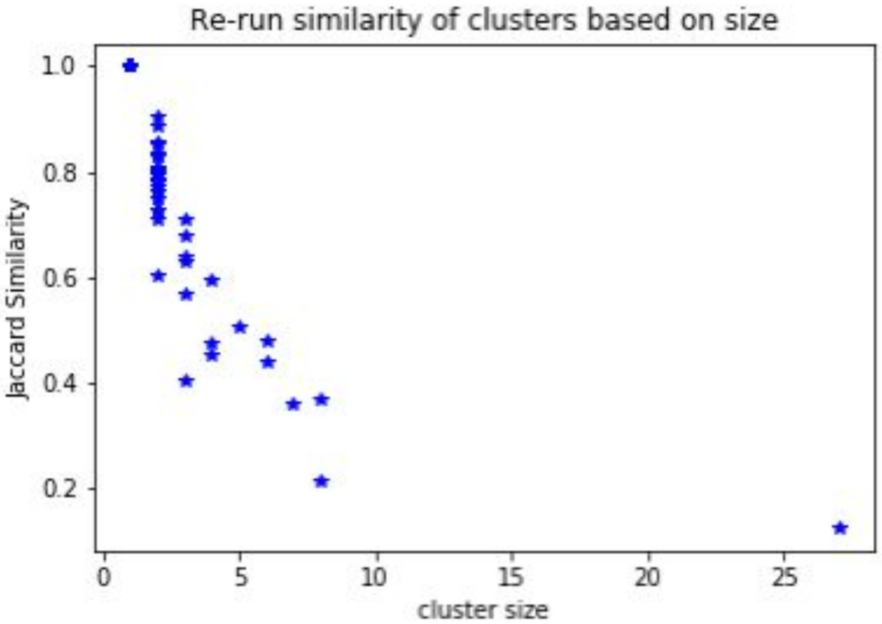
Round 2

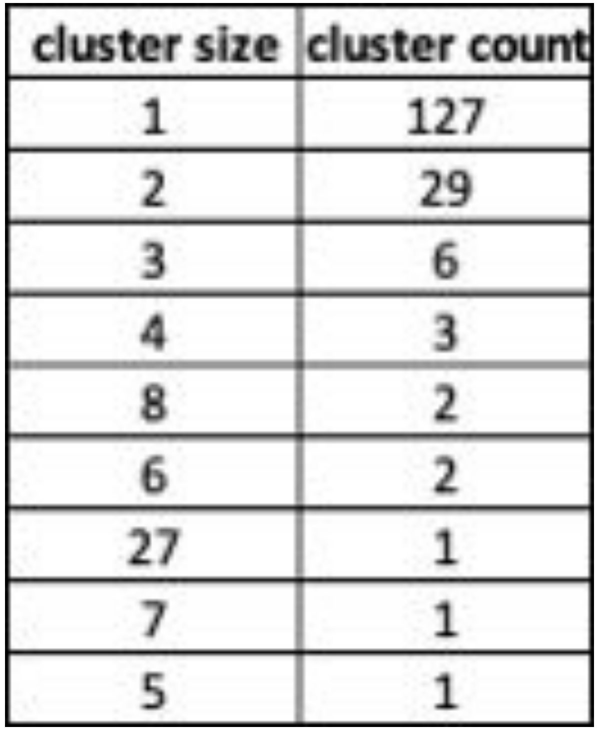




cluster size	cluster count
1	282
2	66
3	27
4	17
5	6
8	3
21	1
18	1
16	1
13	1
12	1
6	1

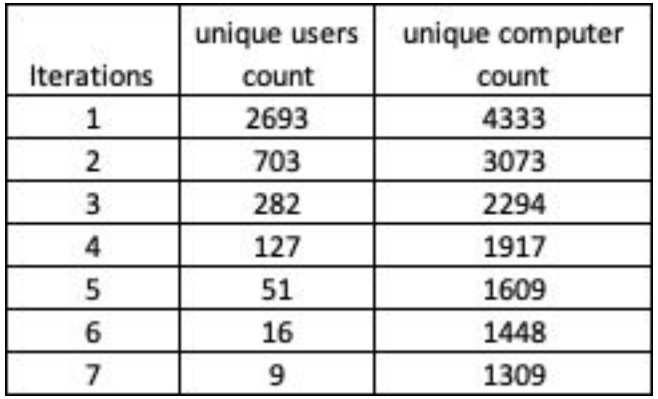
Round 3


Re-run similarity of clusters based on size



cluster size	cluster count
1	127
2	29
3	6
4	3
8	2
6	2
27	1
7	1
5	1

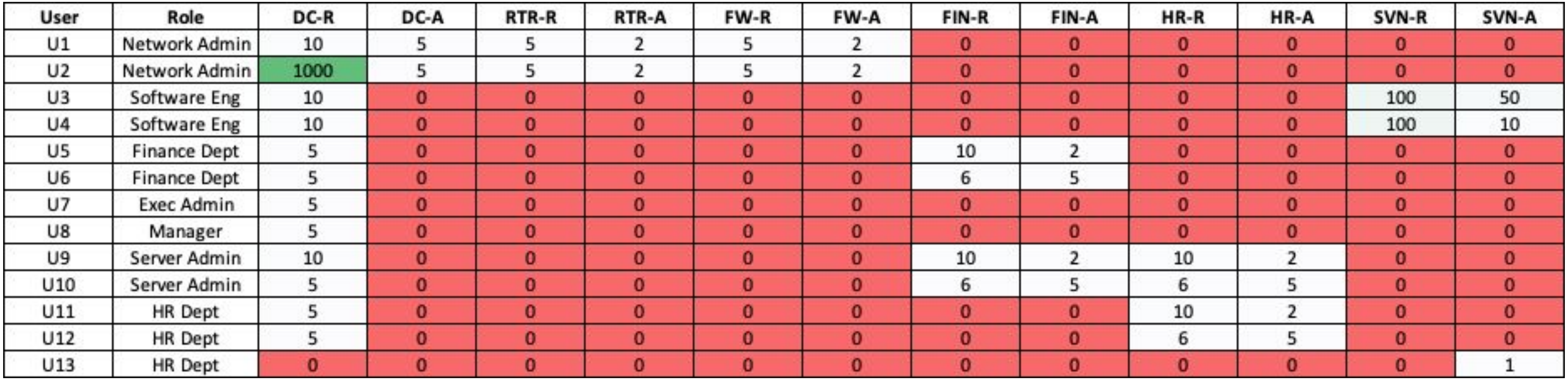
Break down of all rounds



Iterations	unique users count	unique computer count
1	2693	4333
2	703	3073
3	282	2294
4	127	1917
5	51	1609
6	16	1448
7	9	1309

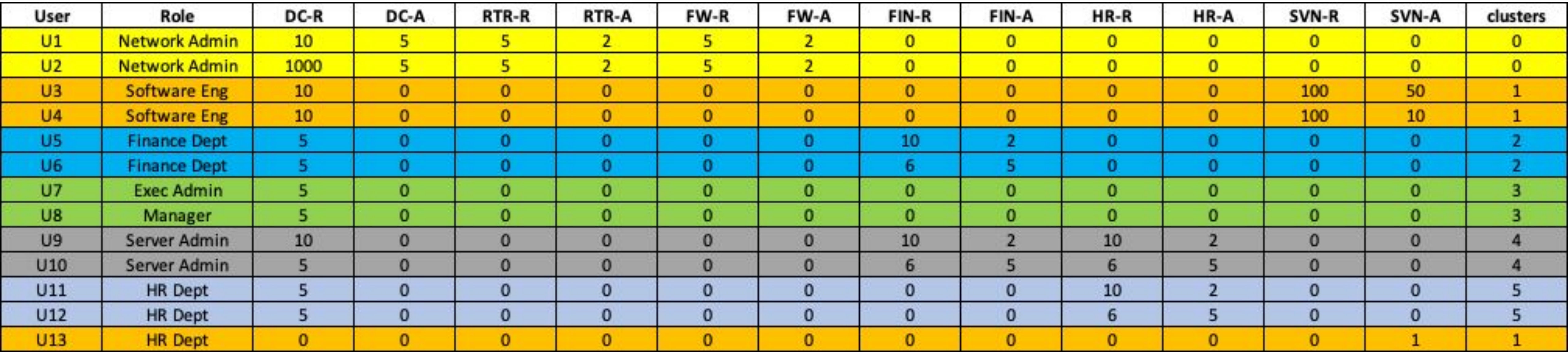
There is a list of 4 users that are left over and not able to be clustered

Toy problem 1



User	Role	DC-R	DC-A	RTR-R	RTR-A	FW-R	FW-A	FIN-R	FIN-A	HR-R	HR-A	SVN-R	SVN-A
U1	Network Admin	10	5	5	2	5	2	0	0	0	0	0	0
U2	Network Admin	1000	5	5	2	5	2	0	0	0	0	0	0
U3	Software Eng	10	0	0	0	0	0	0	0	0	0	100	50
U4	Software Eng	10	0	0	0	0	0	0	0	0	0	100	10
U5	Finance Dept	5	0	0	0	0	0	10	2	0	0	0	0
U6	Finance Dept	5	0	0	0	0	0	6	5	0	0	0	0
U7	Exec Admin	5	0	0	0	0	0	0	0	0	0	0	0
U8	Manager	5	0	0	0	0	0	0	0	0	0	0	0
U9	Server Admin	10	0	0	0	0	0	10	2	10	2	0	0
U10	Server Admin	5	0	0	0	0	0	6	5	6	5	0	0
U11	HR Dept	5	0	0	0	0	0	0	0	10	2	0	0
U12	HR Dept	5	0	0	0	0	0	0	0	6	5	0	0
U13	HR Dept	0	0	0	0	0	0	0	0	0	0	0	1

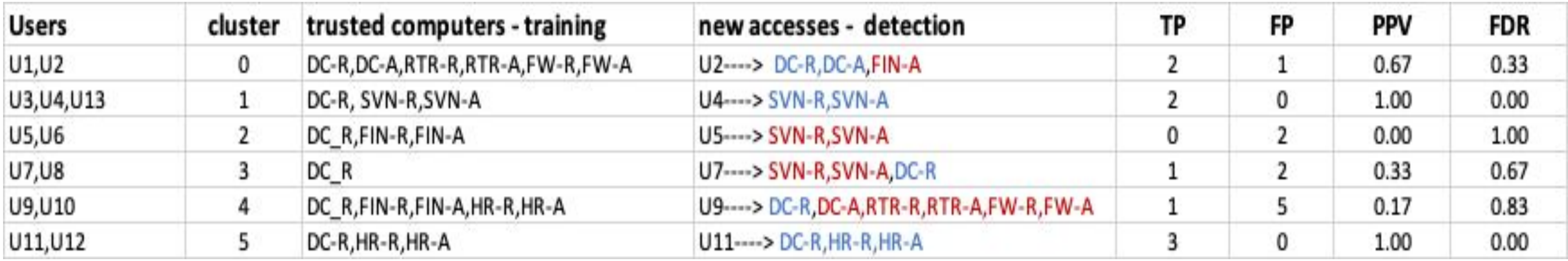
Toy problem 1 – Louvain + KNN clustered



User	Role	DC-R	DC-A	RTR-R	RTR-A	FW-R	FW-A	FIN-R	FIN-A	HR-R	HR-A	SVN-R	SVN-A	clusters
U1	Network Admin	10	5	5	2	5	2	0	0	0	0	0	0	0
U2	Network Admin	1000	5	5	2	5	2	0	0	0	0	0	0	0
U3	Software Eng	10	0	0	0	0	0	0	0	0	0	100	50	1
U4	Software Eng	10	0	0	0	0	0	0	0	0	0	100	10	1
U5	Finance Dept	5	0	0	0	0	0	10	2	0	0	0	0	2
U6	Finance Dept	5	0	0	0	0	0	6	5	0	0	0	0	2
U7	Exec Admin	5	0	0	0	0	0	0	0	0	0	0	0	3
U8	Manager	5	0	0	0	0	0	0	0	0	0	0	0	3
U9	Server Admin	10	0	0	0	0	0	10	2	10	2	0	0	4
U10	Server Admin	5	0	0	0	0	0	6	5	6	5	0	0	4
U11	HR Dept	5	0	0	0	0	0	0	0	10	2	0	0	5
U12	HR Dept	5	0	0	0	0	0	0	0	6	5	0	0	5
U13	HR Dept	0	0	0	0	0	0	0	0	0	0	0	1	1

Calculate accuracy of detection

Cls_{ts} C_d



Users	cluster	trusted computers - training	new accesses - detection	TP	FP	PPV	FDR
U1,U2	0	DC-R,DC-A,RTR-R,RTR-A,FW-R,FW-A	U2----> DC-R,DC-A,FIN-A	2	1	0.67	0.33
U3,U4,U13	1	DC-R, SVN-R,SVN-A	U4----> SVN-R,SVN-A	2	0	1.00	0.00
U5,U6	2	DC_R,FIN-R,FIN-A	U5----> SVN-R,SVN-A	0	2	0.00	1.00
U7,U8	3	DC_R	U7----> SVN-R,SVN-A,DC-R	1	2	0.33	0.67
U9,U10	4	DC_R,FIN-R,FIN-A,HR-R,HR-A	U9----> DC-R,DC-A,RTR-R,RTR-A,FW-R,FW-A	1	5	0.17	0.83
U11,U12	5	DC-R,HR-R,HR-A	U11----> DC-R,HR-R,HR-A	3	0	1.00	0.00

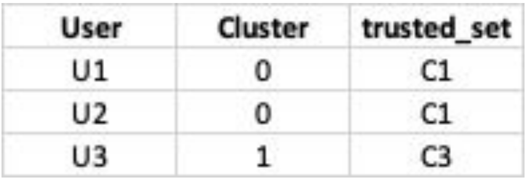
TP – computer accessed in detection and belongs to trusted computer set
FP – computer accessed in detection but does not belong to trusted computer set

PPV (positive predictive value) = TP/(TP + FP) worst = 0, best = 1
FDR (False discovery rate) = 1.0 – PPV OR FP/(TP + FP) best = 0, worst = 1

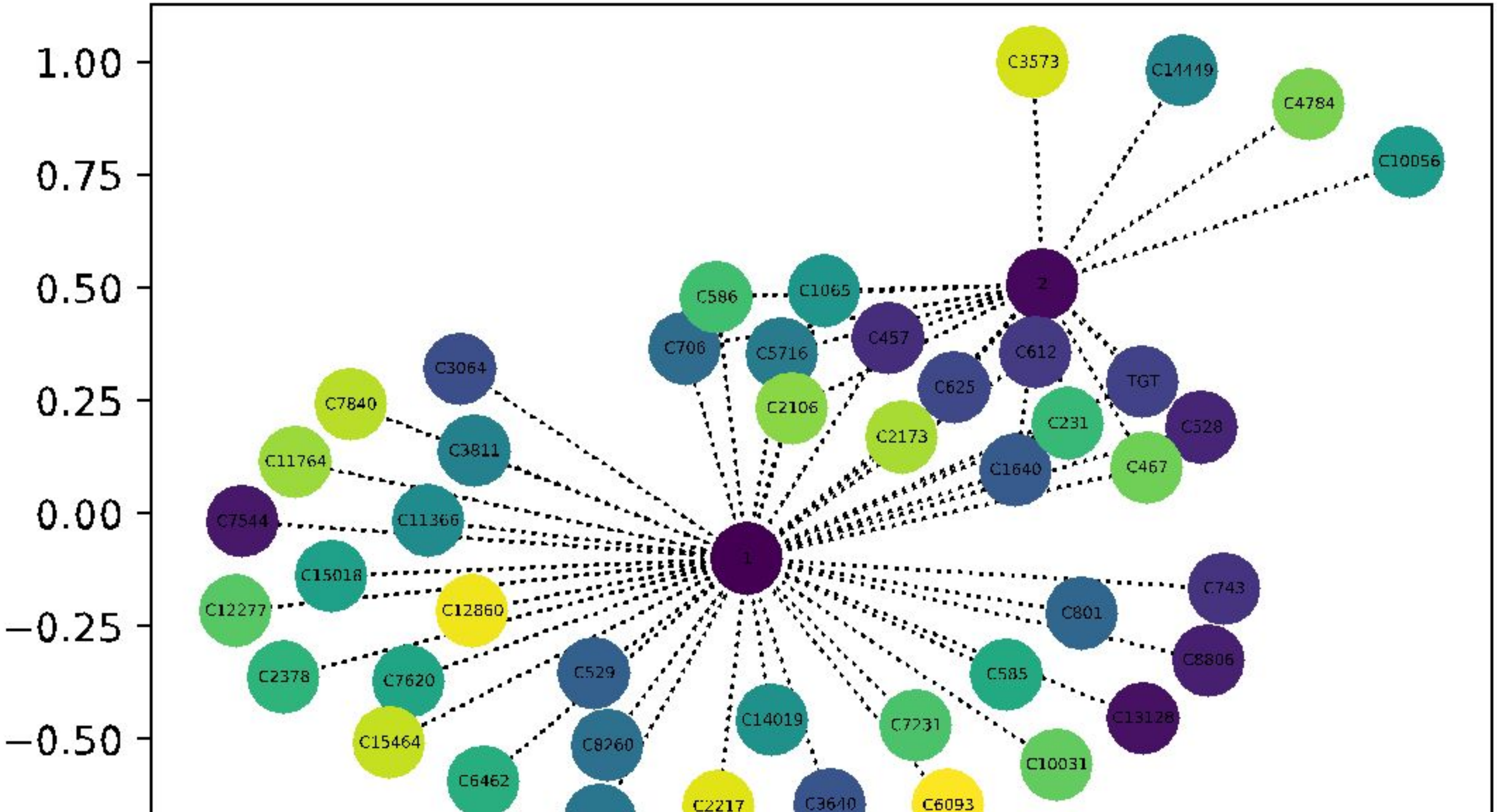
We can use a threshold on PPV or FDR to identify high number False Positives

Detection pseudocode

```
After Training
• For each cluster
    • for all users in cluster
        • combine all computers accessed to form a trusted set $Cls_{ts}$
Detection
• For each user
        • Match user to cluster
        • Get new user computer accesses – $C_d$
        • Use $Cls_{ts}$ and $C_d$ to calculate FDR:
            • Computers common between $Cls_{ts}$ and $C_d$ = TP
            • Computers in $C_d$ and not in $Cls_{ts}$ = FP
            • FDR = FP/(TP + FP)
• Create alert based on FDR
• Need to think of how to handle special cases
```



User	Cluster	trusted_set
U1	0	C1
U2	0	C1
U3	1	C3


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
C3573
C14119
C4784
C10056
2
C586
C1065
C706
C5716
C457
C612
C625
TGT
C2106
C2173
C231
C528
C3064
C7840
C3811
C11764
C1640
C467
C7544
C11366
1
C15018
C743
C12277
C12860
C801
C8806
C529
C2378
C7620
C585
C14019
C7231
C13128
C15464
C8260
C10031
C6462
C2217
C3640
C6093