Case Study BPI 2015 M3

Index

Overview

Dataset : #29 attributes under following levels & categories

- **Event**
  - Activity | concept:name | **activityNameEN** |activityNameNL | **action_code**: Activity Identifier
  - **Resource** : Resource executing the event
  - Monitoring resource : Resource monitoring the event
- **Trace**
  - **Case ID** : Unique case identifier #
  - (case) SUMleges: Cost of each trace
  - (case) last_phase: last phase for each case
  - **(case) caseStatus**: status of the case
  - (case) Includes_subCases: if a case has any subcases
  - (case) Responsible_actor: Resource responsible for case
  - (case) case_type: Type of permit applied

Overview

- **Timestamp**
  - **Complete Timestamp** : Timestamp of execution of an event
  - dateFinished : almost similar to Complete Timestamp
  - dateStop: Not relevant. (couldn’t make any sense)
  - Planned : Planned time of execution of an event
  - dueDate : Due date of execution of an event
- **Others**
  - **Variant** : Cases following similar trace
  - All other 10 attributes are not considered as most of them had missing values.

EDA

Attributes	Count	Comments
Case Id	1409	**Relevant**
Events	59681	**Relevant**
Complete Timestamp	2010.01.01 - 2015.03.05	**Duration of each case**
Variant	1349	**Relevant**
(case) last_phase	27	**Mismatch with activityNameNL**
Action_code | Concept: name	383	**Built new attributes**
activityNameEN	277	**Context Information**

Attributes	Count	Comments
Executing Resources	14	**Relevant**
Responsible_actor	20	**Dropped due to ambiguity**
monitoringResource	22	**Not relevant**
(case) case_type	1 - 557668	**No Insight**
(case) IdofConceptCase	805 (including null)	**No Insight**
(case) caseStatus	2	**Open or Close**
Avg no events per case	43	**Duration & trace length**

Attributes	Count	Comments
(case) parts	94	Not relevant
(case) requestcomplete	2	Not relevant
(case) termName	14	Not relevant
Lifecycle: transition	1	Not relevant
question	540	No Value added
IDofConceptCase | Includes_subCases | SUMleges | caseProcedure | landRegisterID |	40% | 24% | 35% | 87% | 80%	**Missing Data** : dropped

Data Preparation

- In this section we shall discuss:
  - Data preprocessing and cleaning
  - Feature Engineering
  - Attribute selection

- Python, Excel & Disco were used to perform the task.

Preprocesing & cleaning

- The data had logging inconsistencies
  - Timestamp: many had “00:00:00” but had different order. (No changes done)
    - Batch execution ?
    - Automated processing ?
  - Event Ordering: Initially data is ordered based on time, but to be sure, sorted again in terms of order.
    - As a result, now we know the correct duration of each event.
    - Did not bring a major change but more concrete analysis further.
    - Overlaps could be used to find problems in the process flow.
  - CaseStatus: (if lastPhase == activityNameNL) ➔ case should have ended.
    - There are many scenarios when caseType is still open “O”
    - Dropping “G” would lead to loss of information but avoid noise.
    - Conclusion: We consider Closed cases in our analysis.

Feature Engineering

- **Action_code**: Since we can't rely on timestamp values as they happen parallelly, So ➔ had to reorder them.

  New attributes are created

  - **Order :** Last three letters are supposed to mention the order of an event
  - Issues:
    - Created new columns and cleaned a lot of "_" values. (01_BB_1_xxx_Y)
    - String to numeric conversion and
    - Data ordered ➔ first by **Case ID** ➔ and then **Order**.

  - **Phase** :
    - **01_H00FD_1xx** ➔ It is informed the activity belongs to Phase 1
    - Observed a total of 10 phases from (0-9) and their frequency distribution is shown below.

Phase	Count	Percentage
0	22973	38.5
4	10772	18.0
5	9859	16.5
1	6937	11.6
3	4670	7.8
2	4105	6.9
8	188	0.3
7	107	0.2
6	33	0.1
UVO	22	0.0
9	15	0.0
	59681	100.0

Phase level frequency

- Summarizing for all 277 activities would be too fine grained.
- For Simplicity, looked at main stages/phases for all the activities (0-9)
- Phase [0,1,4,5] contains 85 % of all activities in log
- There was a label ‘UOV’ in the action code, but they represent the subprocess count.
- delete these records as we do not have any details about the order of these activities

Feature Engineering

- **Activity Name:**
  - Many to one relationship : Some activity names have more than one action code (383 ➔ 277)
  - Difficult to make a process model with 383 activities.
  - Suggestion : Aggregation by abstracting to a higher phase level
  - We find 53 distinct activityNameEN with the word “phase”
  - Issue: These names can’t be related to action_code. (Reason activityNameNL might make sense).
  - **Conclusion**: Keep things simple and use action_code to identify main events.

- **Completeness:**
  - Ongoing cases ➔ ! process discovery of case variants.
  - Three attributes:
    - requestComplete | caseStatus | endDate
    - Nothing gave an implicit conclusion to decide if the case has actually ended.
    - As discussed above, we only consider caseStatus.

Feature Engineering

- **Main/Sub-process**: This would be the coarse grained indicator of each event.
  - **01_H00FD_1xx** : Split out the central part and make a frequency table. Described below.

- **Timestamp**:
  - Complete:timestamp | dateFinished | planned | dateStop | dueDate - *Event Level*
  - startDate | endDate | endDatePlanned - *Case Level*
  - Since most of them were **missing** → “NA”
  - Timestamp of last event does’t coincide with endDate
  - **Conclusion**: they are not reliable, We only consider **Complete:timestamp**

- Now we have two levels on which model could be developed
  - **Phase**: Range(0,9)
  - **Subprocess** : 18 values (Table below)
  - **Issue**: We do not have any context information in this approach.

subprocess	count	Percentage
HOOFD	45557	76.3%
AWB45	3642	6.1%
AH	2612	4.4%
BPT	1534	2.6%
VD	1006	1.7%
UOV	926	1.6%
GBH	909	1.5%
DRZ	884	1.5%
AP	668	1.1%
EIND	585	1.0%
CRD	501	0.8%
OPS	343	0.6%
VRIJ	242	0.4%
BB	135	0.2%
NGV	119	0.2%
OLO	13	0.0%
LGSV	4	0.0%
LGSD	1	0.0%
	59681	100.0

Main/Subprocess Freq

- **No relation of each subprocess with the context information.**
  - **Reason:** believe it would match with activityNameNL if we do some word cloud analysis on each activity name & then try to find out some relation.

- **Further Model development will be described**
  - **With subprocess code**
  - **Phase number 0-9**

Attrubute Selection

After dropping caseType == “G” and Phase = “UOV” (22 records). We take following attributes for our analysis

- **Case ID :** **# 1328**
- **Complete Timestamp:** **01:01:2010 - 05:03:2015**
- **Phase:** **# 9**
- **Subprocess:** **# 18**
- **ActionCode :** **# 374**
- **activityNameEN:** **# 276**
- **Order:** **# 128**
- **Resource:** **# 14**
- **caseStatus:** **# 1** Note: We will only consider “O” henceforth
- **Variant:** **# 1285**

Statistics

Observation	Value	Comments
Events	57467	Data loss of 3.7%
Case duration: Mean | Median	39d | 62.6 d	60 % cases → takes >= 40d ~28% cases → takes >=60d 22% cases → 40< bw <60
Variants	~600/1285	Represents our 75% of model behaviour
Phase	0 | 1 | 4 | 5	Cumulates for ~85 % of events
Subprocess codes	HOOFD | AWB45 | AH | BPT	~ 90 % events are in this phases
Active case / day	Range (14 - 18)	From 10-2010 → observe an increase. However with time the figure has slowly decreased

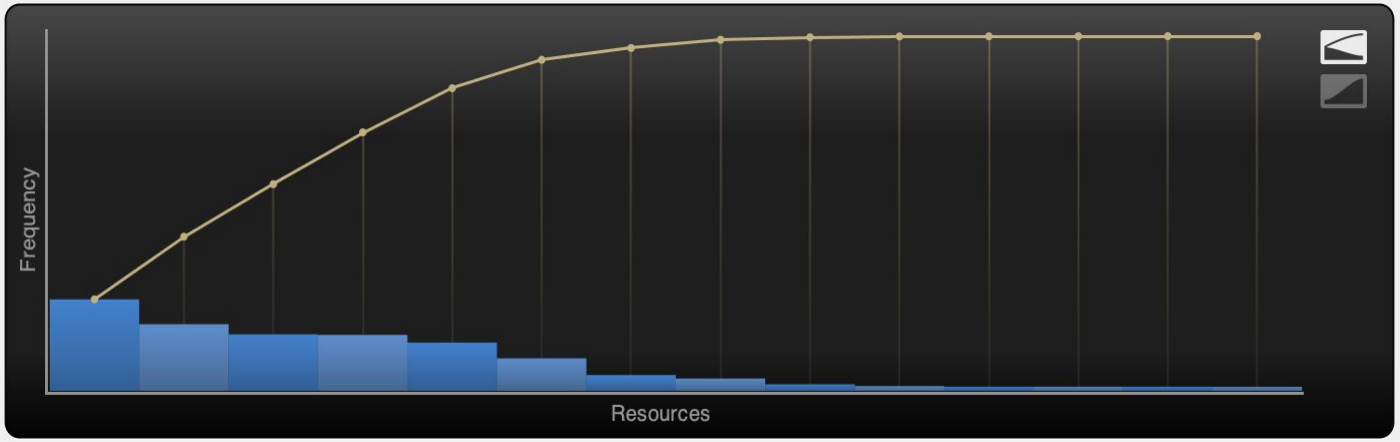
Resource Roles

First step would be to limit an attribute.

- **Resource : 14 members - Event Level**
- Monitoring Resource :22 members - responsible for cases.
    - Relevance when some Case Ids involve multiple municipalities.
    - So, not relevant to our analysis.
- Responsible Actor : 20 unique members who monitor at Trace level

In order to find the activities performed by each resource, a frequency table is shown below.

- Mean - 4106 & Median - 1645. Distribution is heavily right skewed.
- We conclude 5 resources contribute in 86% of the activities.
- **Conclusion**: Most of the workload is on 5 particular resources.



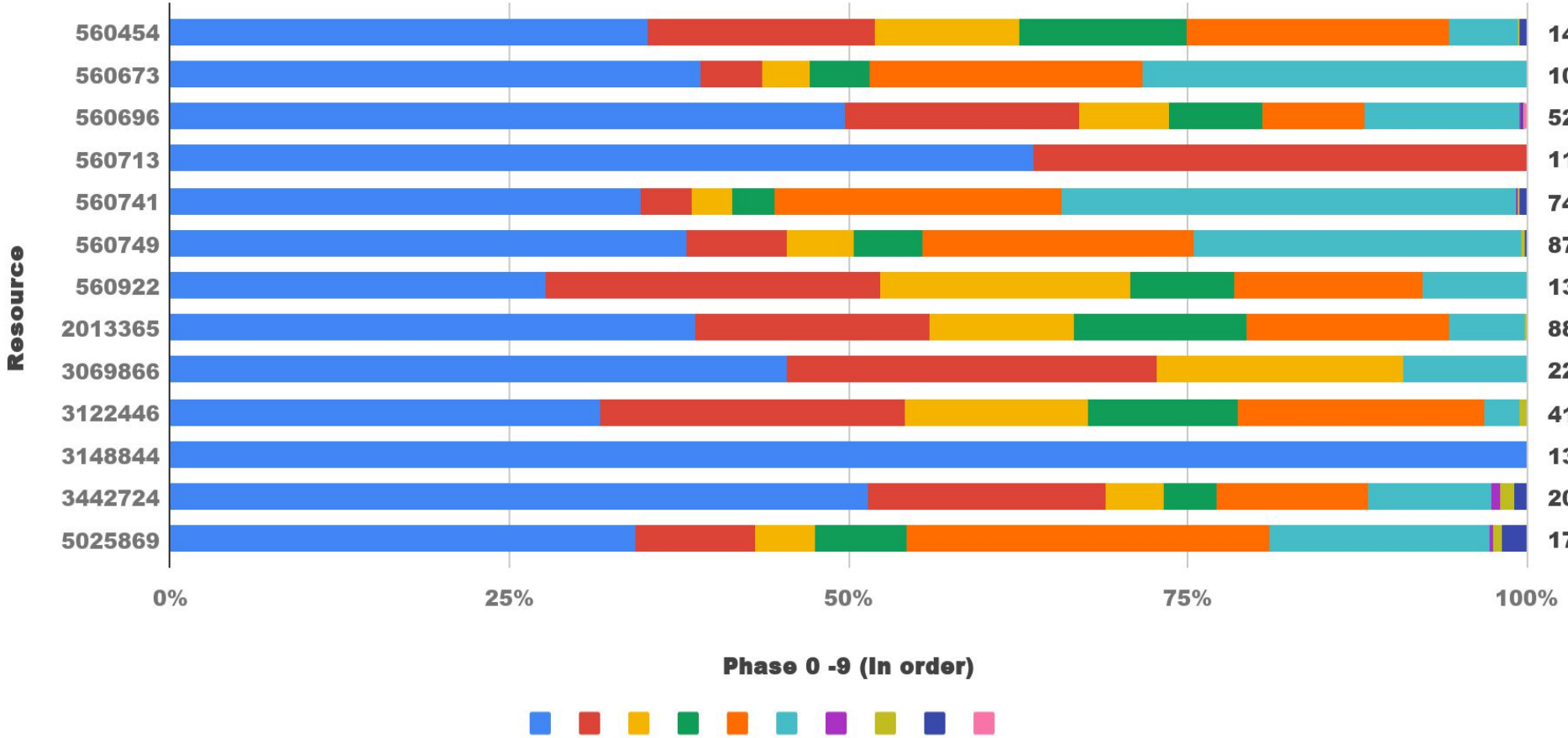
Resources	14
Minimal frequency	2
Median frequency	1,645
Mean frequency	4,106.29
Maximal frequency	14,339
Frequency std. deviation	4,812.59

All resources (14) | First in case (10) | Last in case (10)

Resource	▲ Frequency	Relative frequency
560454	14,339	24.94 %
560673	10,269	17.86 %
2013365	8,609	14.98 %
560749	8,493	14.77 %
560741	7,241	12.6 %
560696	4,652	8.09 %
3442724	1,940	3.37 %
5025869	1,350	2.35 %
3122446	417	0.73 %
560922	130	0.23 %
3069866	22	0.04 %
3148844	13	0.02 %
560713	11	0.02 %
6	2	0 %

Resource	0	1	2	3	4	5
560454	35.2	16.6	10.7	12.3	19.4	5.0
560673	39.0	4.6	3.4	4.5	20.1	28.2
560696	49.8	17.2	6.7	6.8	7.6	11.4
560713	63.6	36.4	0.0	0.0	0.0	0.0
560741	34.7	3.8	2.9	3.1	21.2	33.5
560749	38.0	7.5	4.8	5.1	20.1	24.1
560922	27.7	24.6	18.5	7.7	13.8	7.7
2013365	38.7	17.3	10.6	12.7	14.9	5.6
3069866	45.5	27.3	18.2	0.0	0.0	9.1
3122446	31.7	22.5	13.4	11.0	18.2	2.6
3148844	100.0	0.0	0.0	0.0	0.0	0.0
3442724	51.4	17.5	4.3	3.9	11.2	9.1
5025869	34.2	8.9	4.5	6.7	26.8	16.3

Phase Level Frequency



Resource Roles

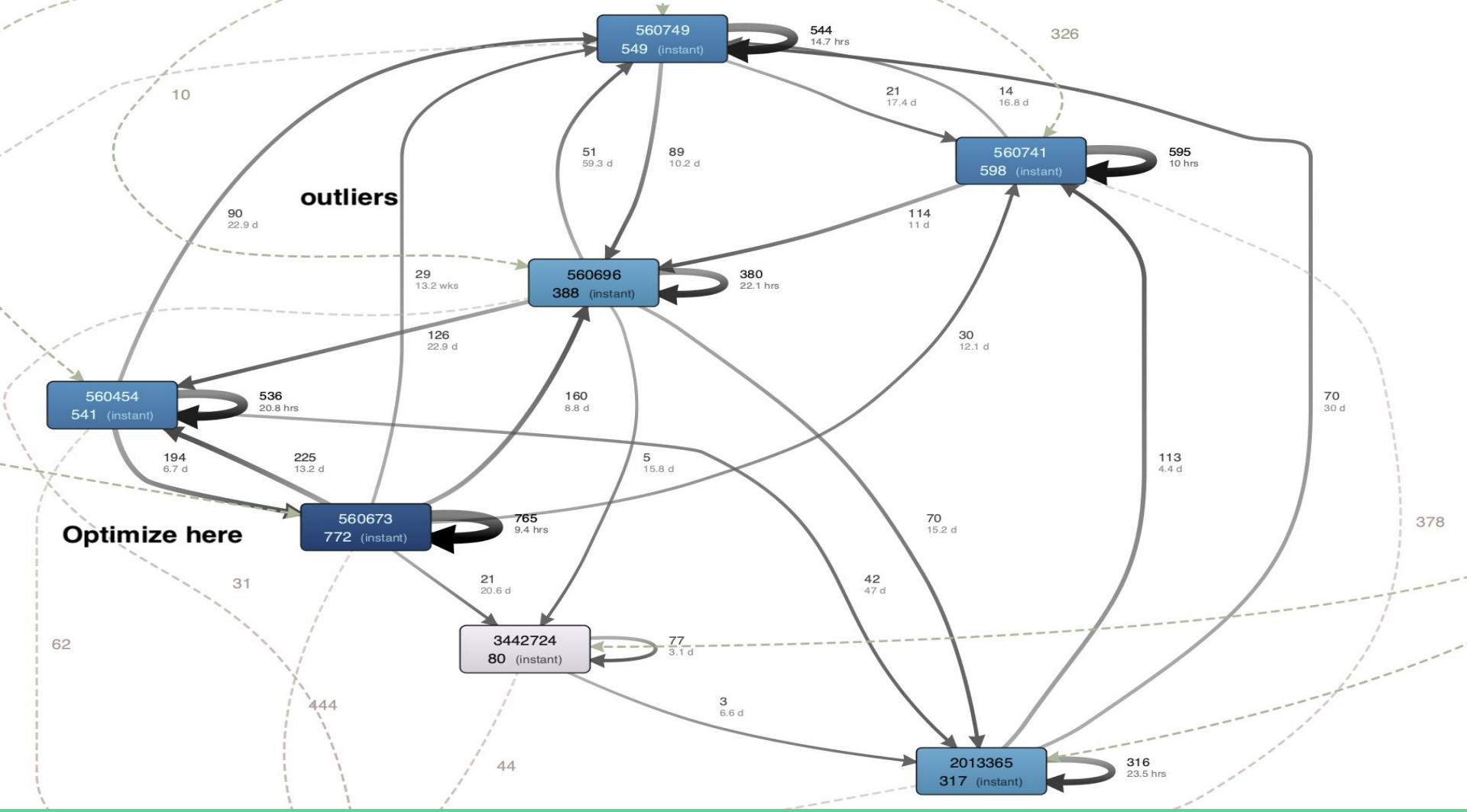
- From the log data, it is not clear how roles are assigned per resource. However, there are 2 cases
  - 6 : Involved in Phase 2
  - 3148844: Involved in Phase 0
- Multiple resources perform different roles. Eg: Resource 560654 has interchanging roles.
  - Resource ➔ 25%
  - Responsible actor ➔ 31%
  - Monitoring resource ➔ 27 %
- Assumption: Could this be due to lack of human resources?
- Comparison of resource roles:
  - All 14 people in Resource ➔ Monitoring resource
  - 8/14 people in Resource ➔ Responsible actor.
- Conclusion: The roles of Resources seem to be interchangeable with Monitoring & Responsible actor.

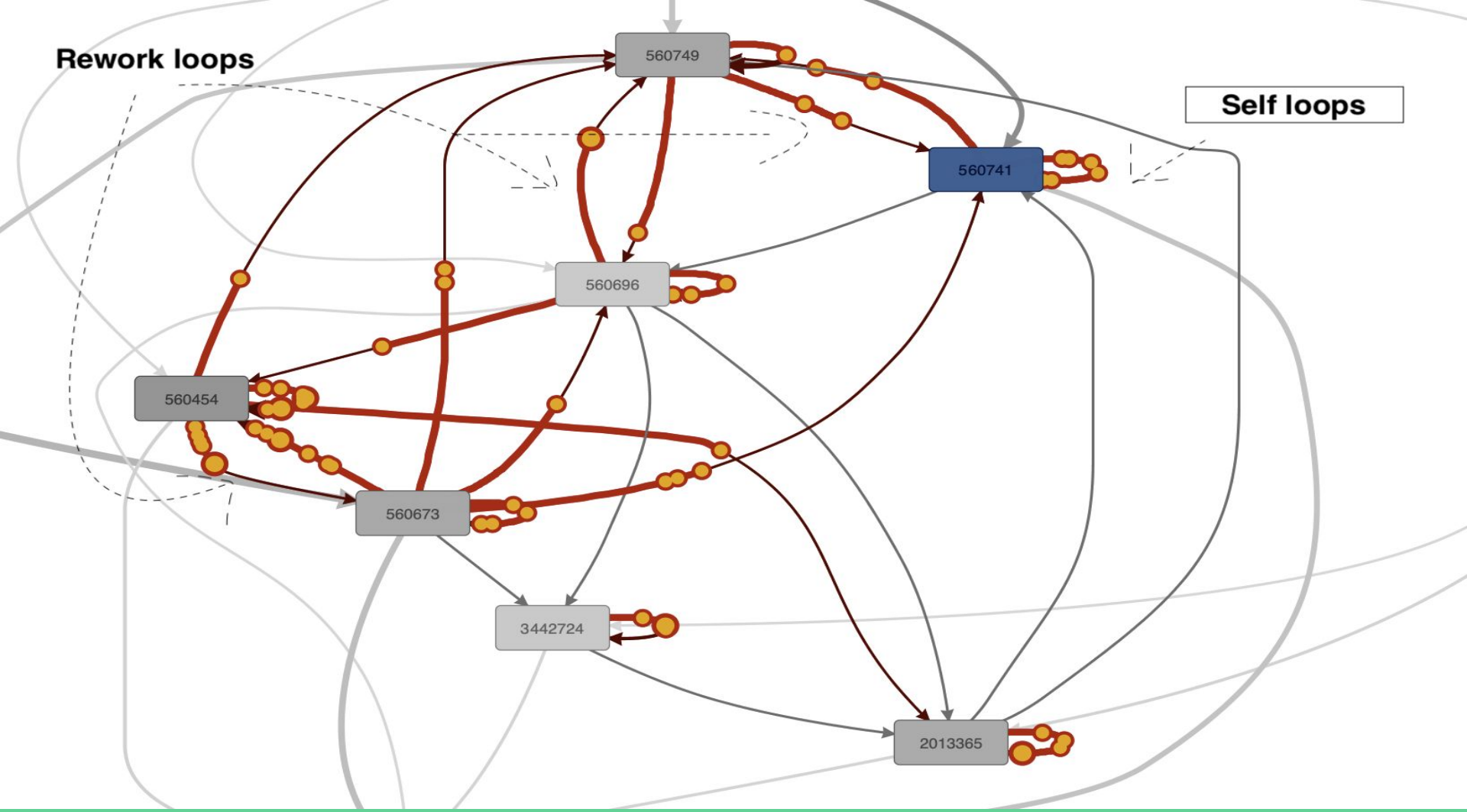
Organisational Structure

- **It is often that a case takes longer due to some resource having huge workload.**
- **We have already identified such resources above and further improvements could be done.**

Observation Parameters : Activity → 50% Path → 90%

- **Many resources take a much higher workload than others (as indicated in the figure below).**
  - **560673 | 560741 | 560749 | 560454 | 560696 | 2013365**
- **Therefore, one of the improvement could be to manage the workload between the resources performing the same role evenly. (**Refer pg no 19**)**
- **Re-do work loops among**
  - **560454 & 560673 → ~ 200 cases**
  - **560749 & 560696 → ~ 60-90 cases**

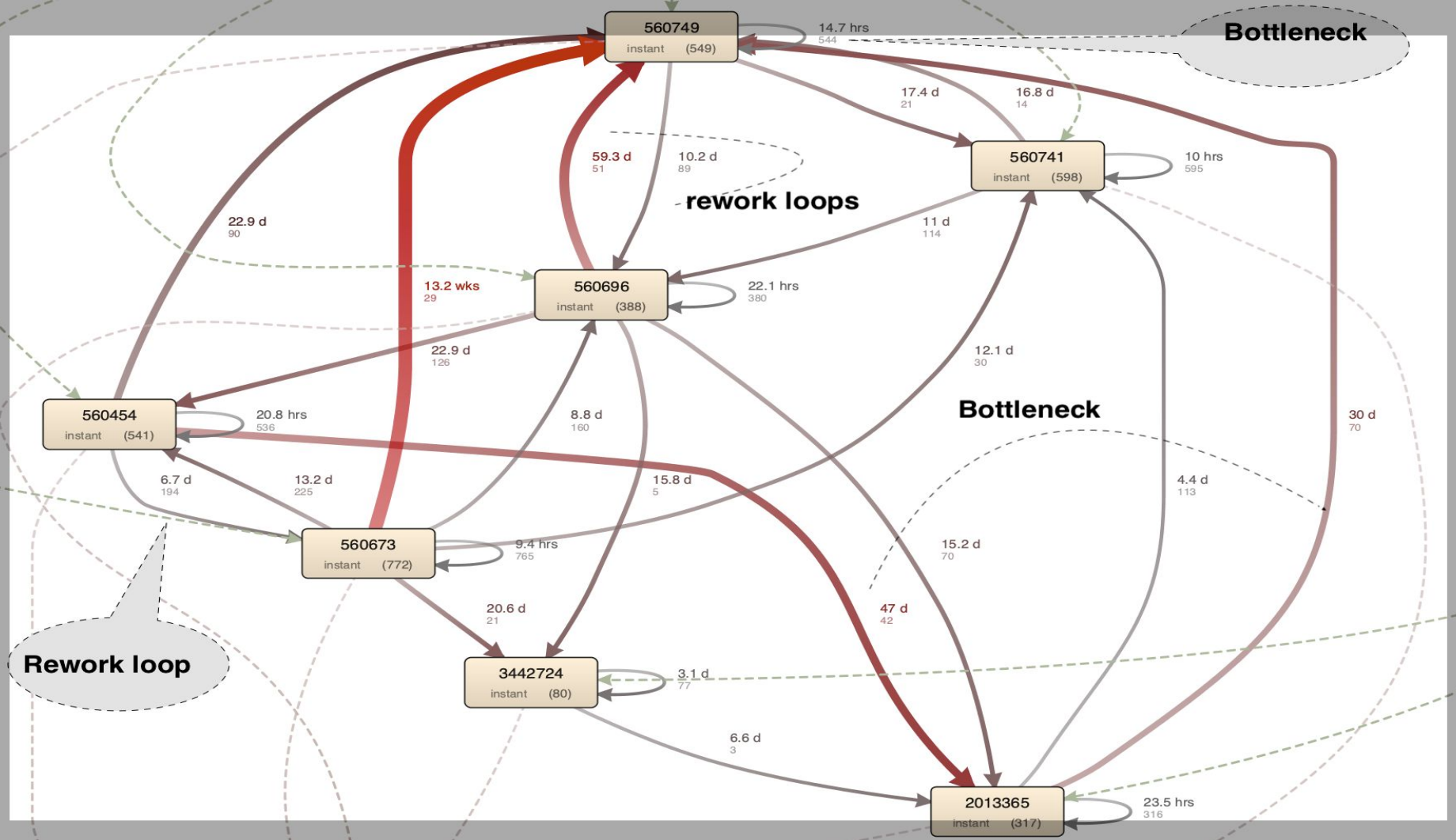

560749
549 (instant)
544
14.7 hrs
326
10
21
17.4 d
14
16.8 d
560741
598 (instant)
595
10 hrs
51
59.3 d
89
10.2 d
outliers
90
22.9 d
114
11 d
29
13.2 wks
560696
388 (instant)
380
22.1 hrs
126
22.9 d
30
12.1 d
560454
541 (instant)
536
20.8 hrs
160
8.8 d
70
30 d
194
6.7 d
225
13.2 d
5
15.8 d
113
4.4 d
Optimize here
560673
772 (instant)
765
9.4 hrs
70
15.2 d
378
31
21
20.6 d
42
47 d
62
3442724
80 (instant)
77
3.1 d
444
3
6.6 d
44
2013365
317 (instant)
316
23.5 hrs


Rework loops
Self loops
560749
560741
560696
560454
560673
3442724
2013365

Organisational Structure

Improvements

- **Self loops :**
  - Tasks are processed periodically, but the impact is overall.
  - Kept on low Priority, that they queue heavily
  - Resources are overloaded that leads to FIFO - (distribute tasks among free resources)
- **High Impact Areas:**
  - **Rework loop** b/w 560454 & 560673 would free 200 cases and save ~ 12 days waiting time
  - 560749 | 560741 : receive **majority incoming cases** and almost all redo work. So their work is an important area to understand domain outside log data. **(Errors or some feedback/approval)**
- **Delegate the rebound tasks to new resources instead of overloaded employees.** Pg20
- **Divide cases based on “currently available” resources; and ! on previous experience.**
  - The latter resources keep waiting for other initial resources to pass on the case, who happen to work more than their capacity. **(~6 do not actively participate in workload)**


560749
instant (549)
14.7 hrs
544
Bottleneck
17.4 d
21
16.8 d
14
560741
instant (598)
10 hrs
595
59.3 d
51
10.2 d
89
rework loops
22.9 d
90
11 d
114
560696
instant (388)
22.1 hrs
380
13.2 wks
29
22.9 d
126
12.1 d
30
560454
instant (541)
20.8 hrs
536
8.8 d
160
Bottleneck
30 d
70
6.7 d
194
13.2 d
225
15.8 d
5
4.4 d
113
560673
instant (772)
9.4 hrs
765
15.2 d
70
20.6 d
21
47 d
42
Rework loop
3442724
instant (80)
3.1 d
77
6.6 d
3
2013365
instant (317)
23.5 hrs
316

Outsourcing

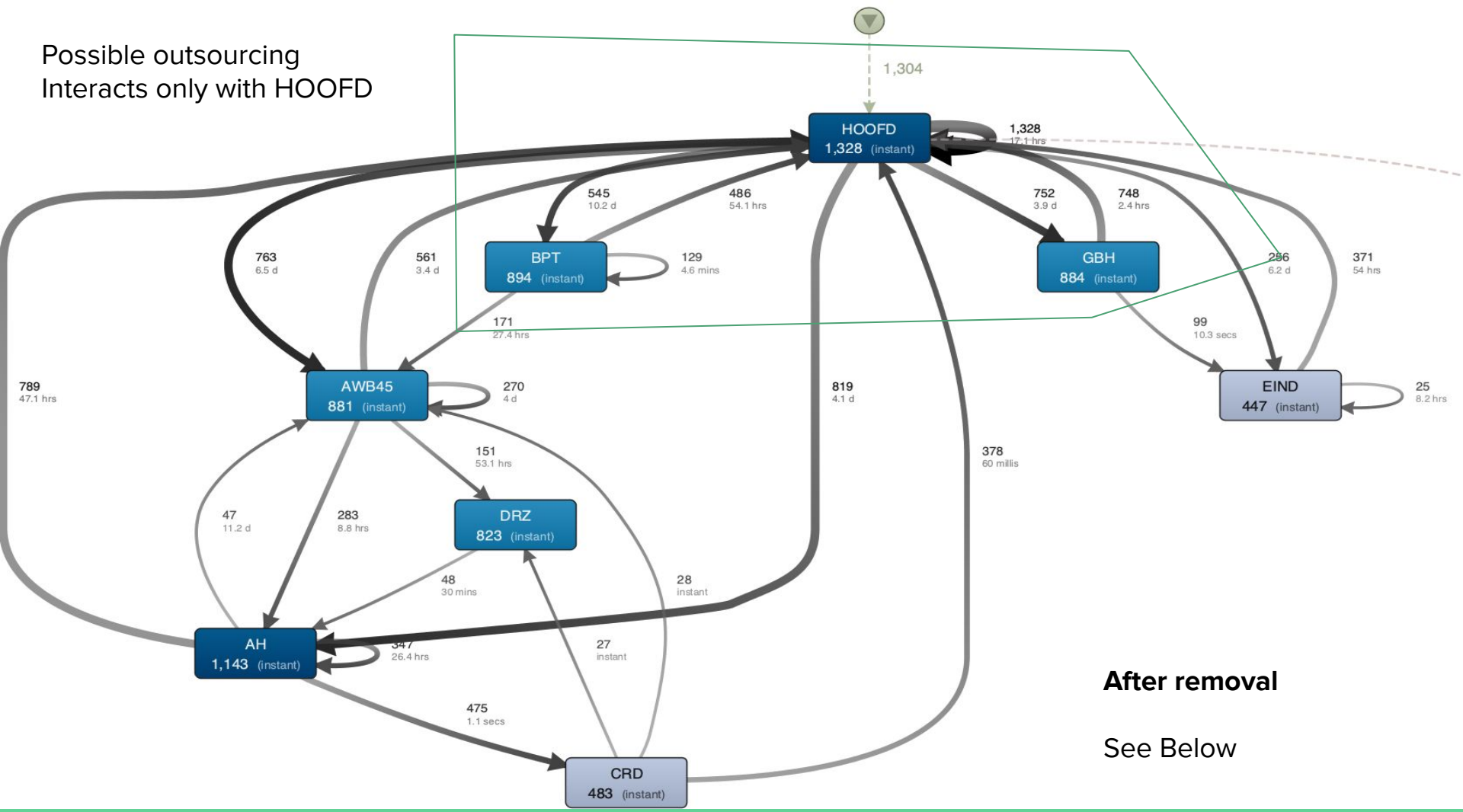
- **Importance: Cost Reduction (Preserve important activities & support from third party)**
- **Better task allocation to the primary resources and not overloaded.**
- **First look on the most frequent subprocess to outsource.** Refer Pg no 13
- **HOOFD → is a main process & ! a good idea to outsource.**
- **Suggestions:**

AWB45	3642	6.1%
AH	2612	4.4%
BPT	1534	2.6%
VD	1006	1.7%
UOV	926	1.6%
GBH	909	1.5%
DRZ	884	1.5%

- **Outsource those which are**

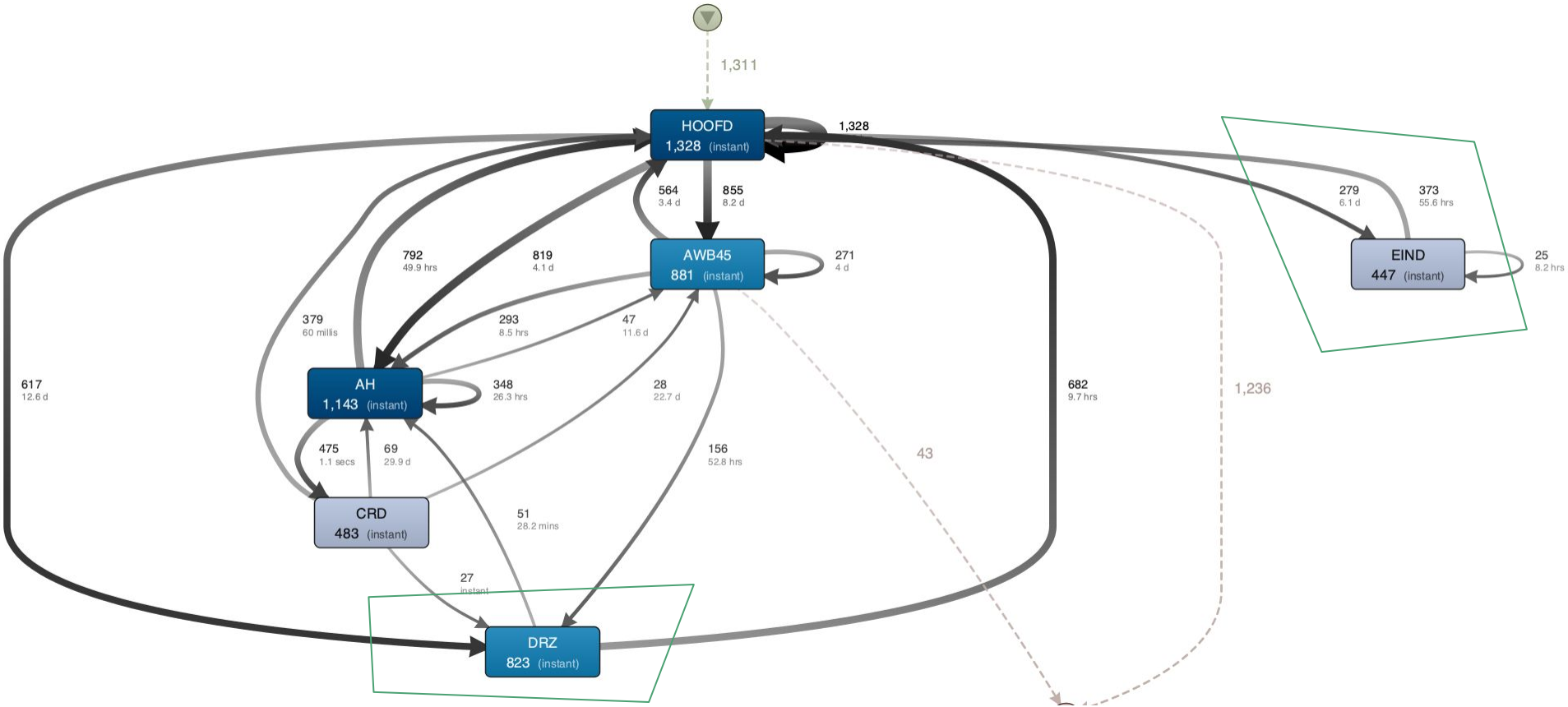
1. **difficult to be Managed**
2. **Doesn't involve loops with other subprocess**
3. **Directly connected to the main process**

Possible outsourcing
Interacts only with HOOFD

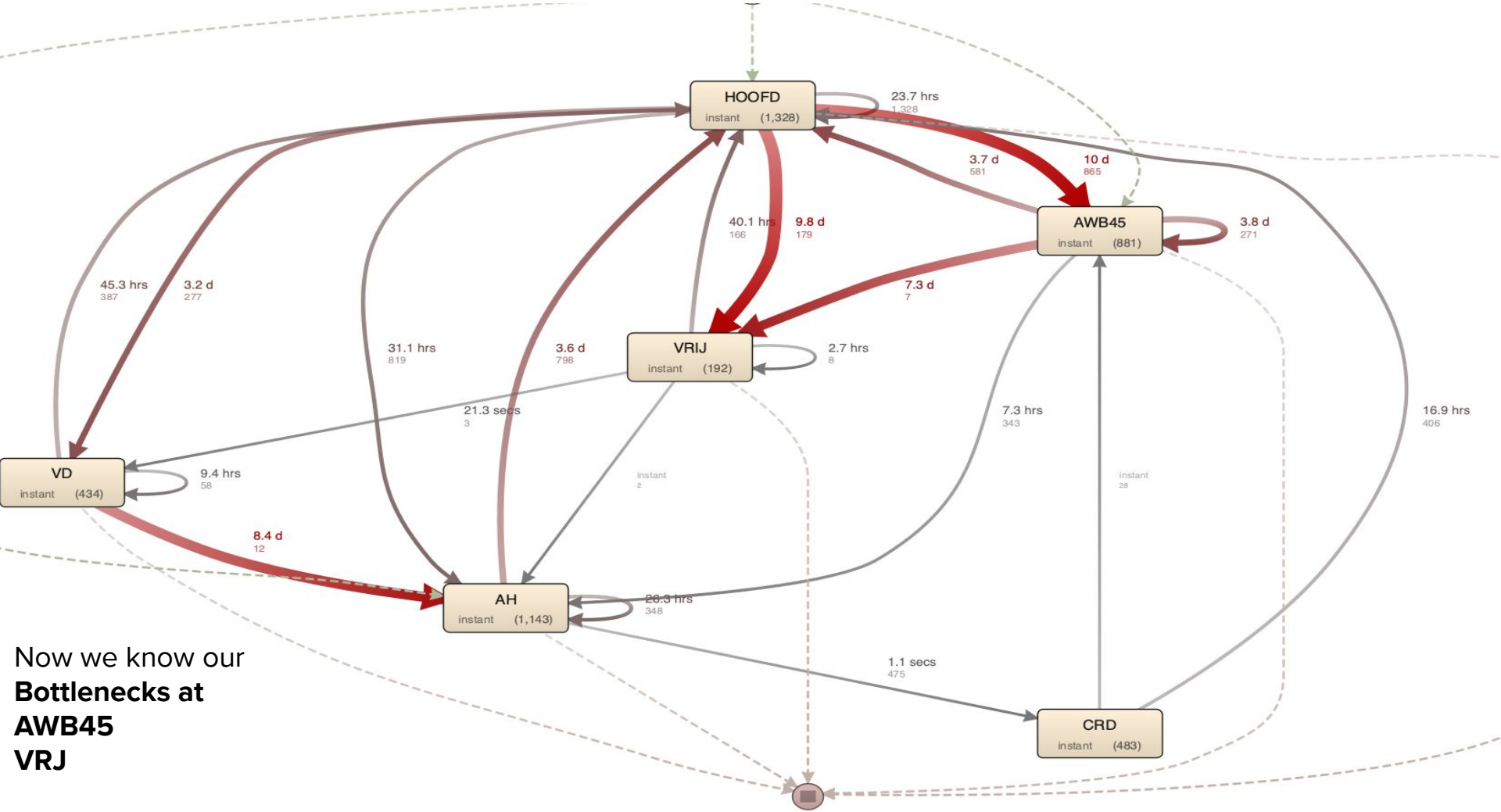


After removal

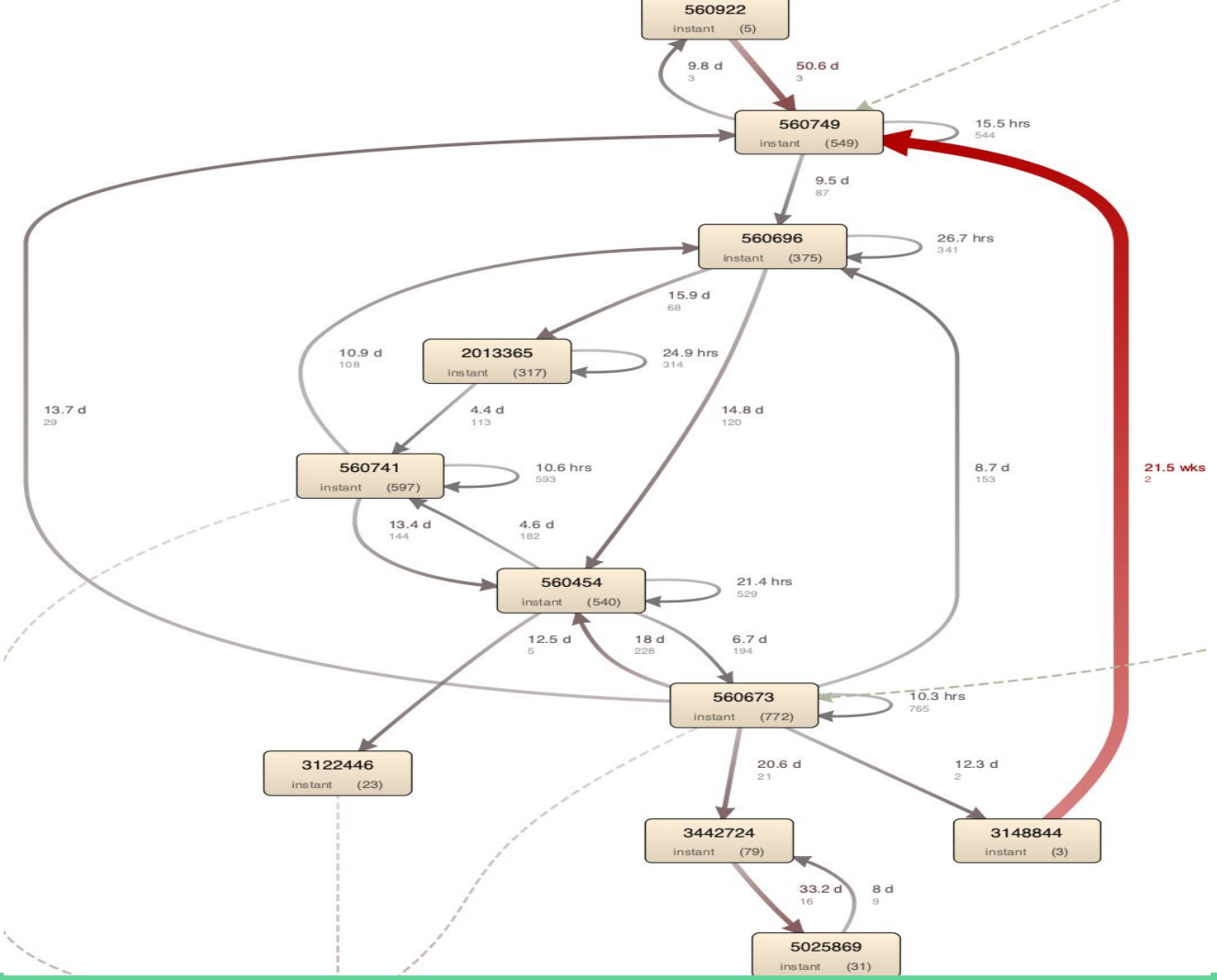
See Below



DRZ : ~650 cases & mean waiting time (12.6d + 9.7 hrs)
EIND : ~ 320 cases & mean waiting time (6.1 d + 55 hrs)



Now we know our
Bottlenecks at
AWB45
VRJ



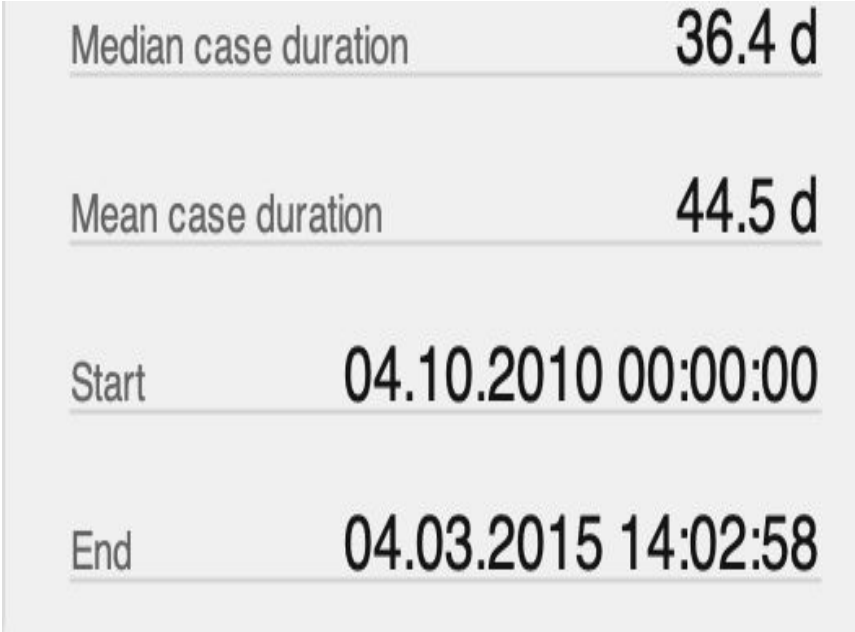
Improvements

Resources are no longer Overloaded.

2 cases which take 21 weeks but they might be an outlier.

Throughput times

- Large difference in mean & median values distribute the data in a **right skew.**
- **Outliers** presence is a huge factor for average 62 days/case.
- As discussed above
    - 60 % cases → takes >= 40 days
    - ~28% cases → takes >=60 days
    - 22% cases → 40< bw <60 days.
- If we look at 95% cases we observe following stats
    - Mean duration shifts from 62 → 46
    - In 2010 until October process took longer Duration and it doesn't help us build a good Process model.



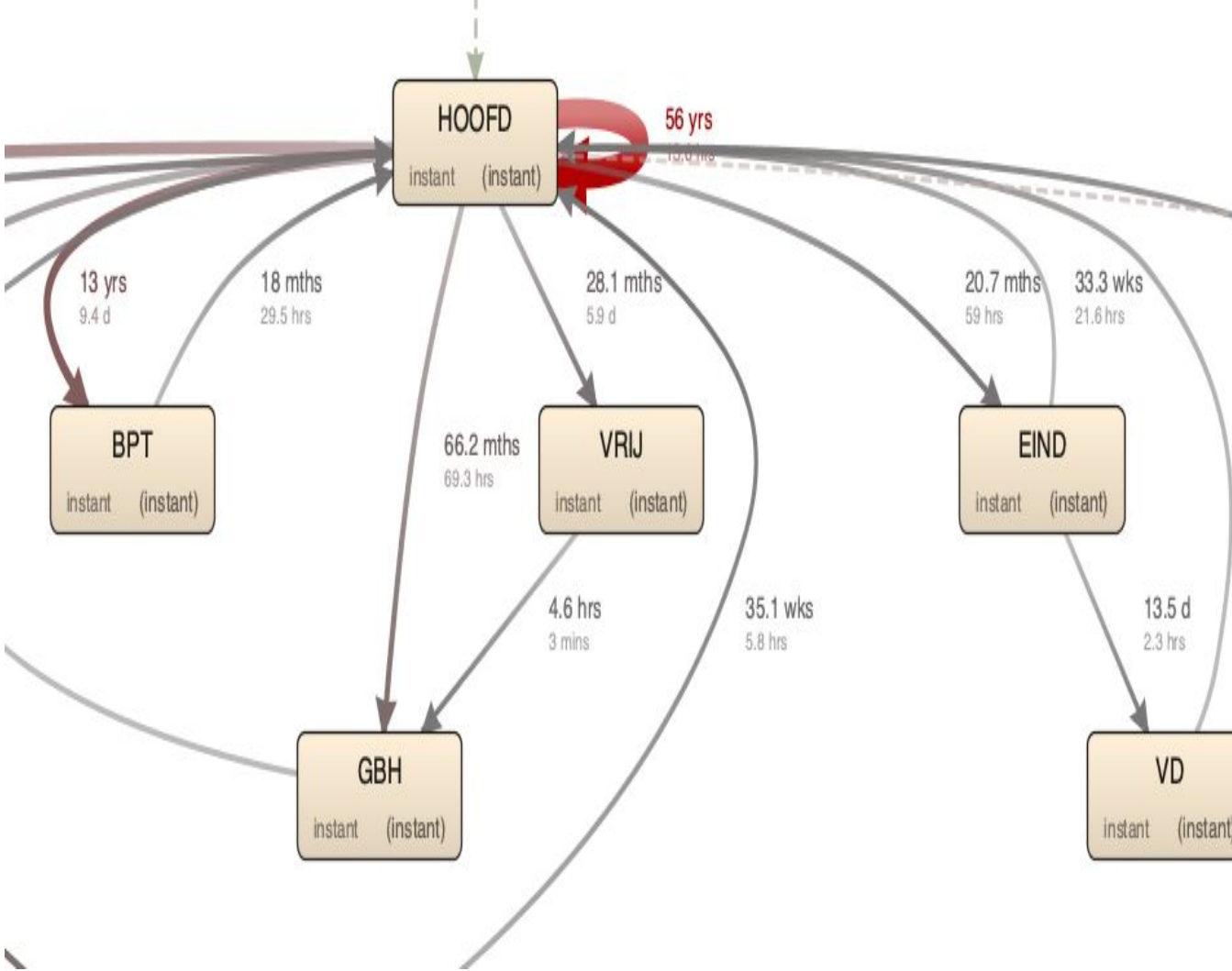
Throughput times

- **Performance Analysis**

For the Action Code improved drastically on removing these 5 % cases.

The most clear bottleneck we observe in HOOFD.

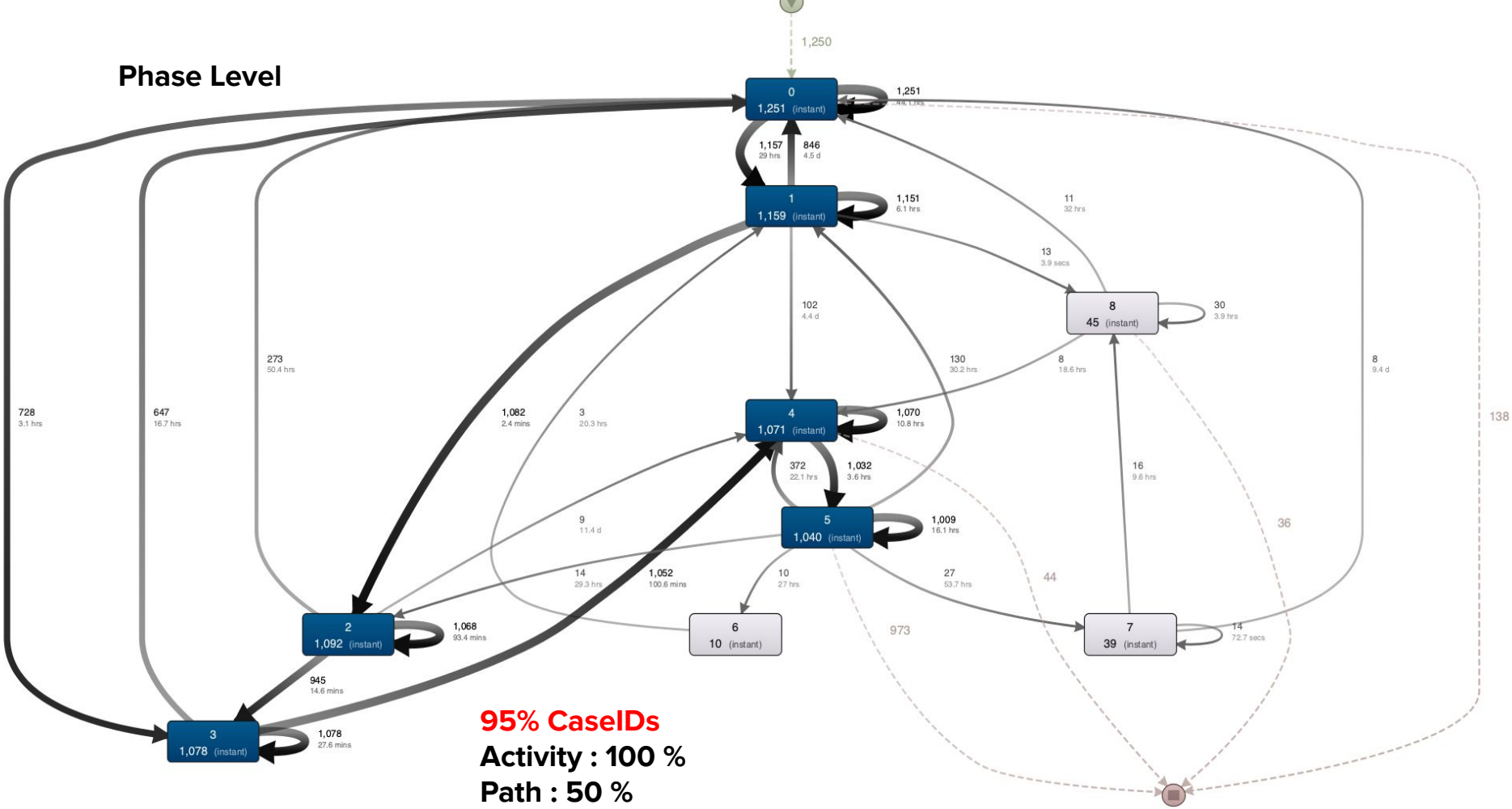
Which Shall require further drill down.



Model 1

- We have two levels of abstraction
    - Action Code level
    - Phase level
- Model would be best represented in absence of outliers and on clean data

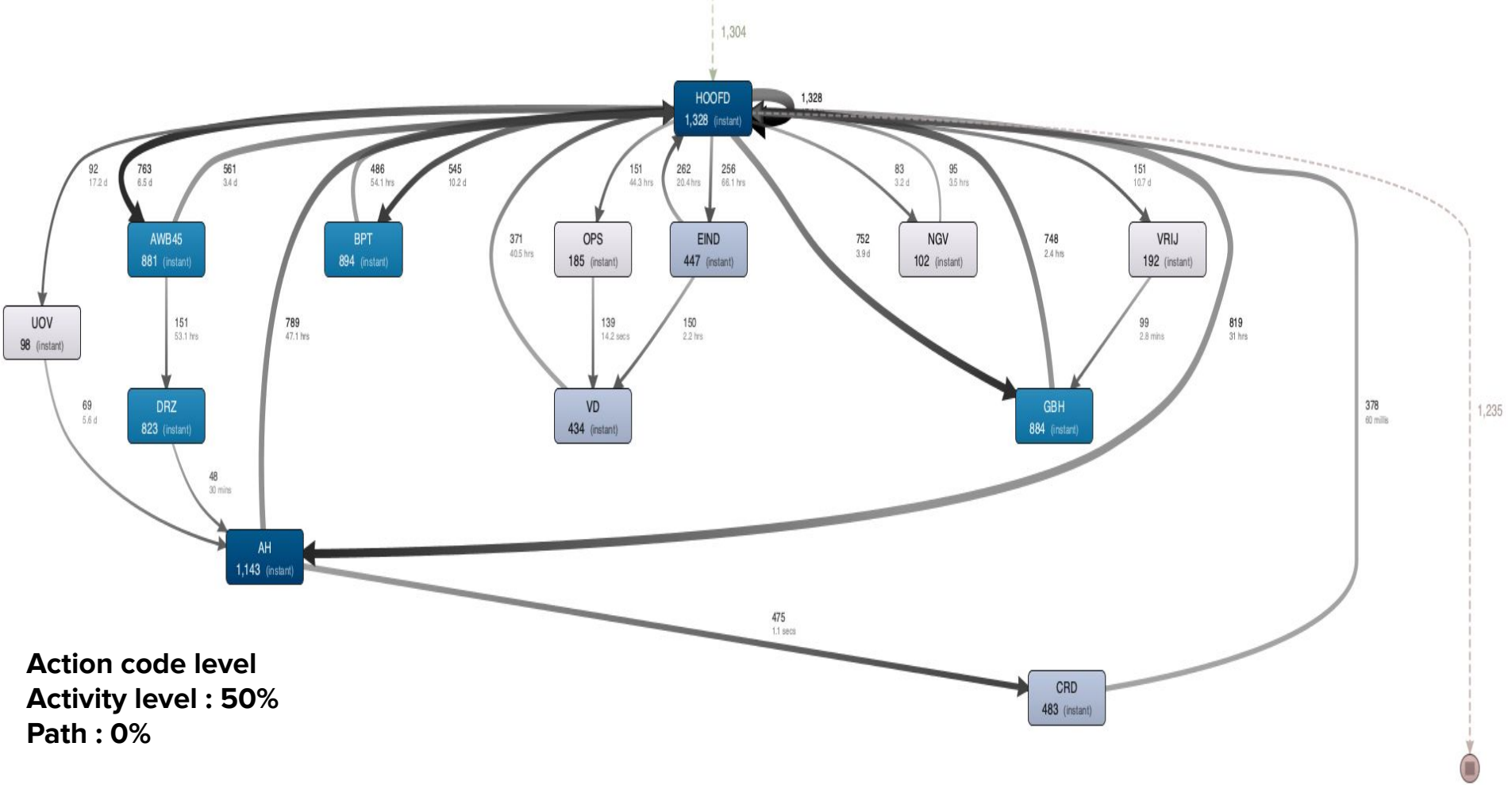
Phase Level



95% CaseIDs

Activity : 100 %

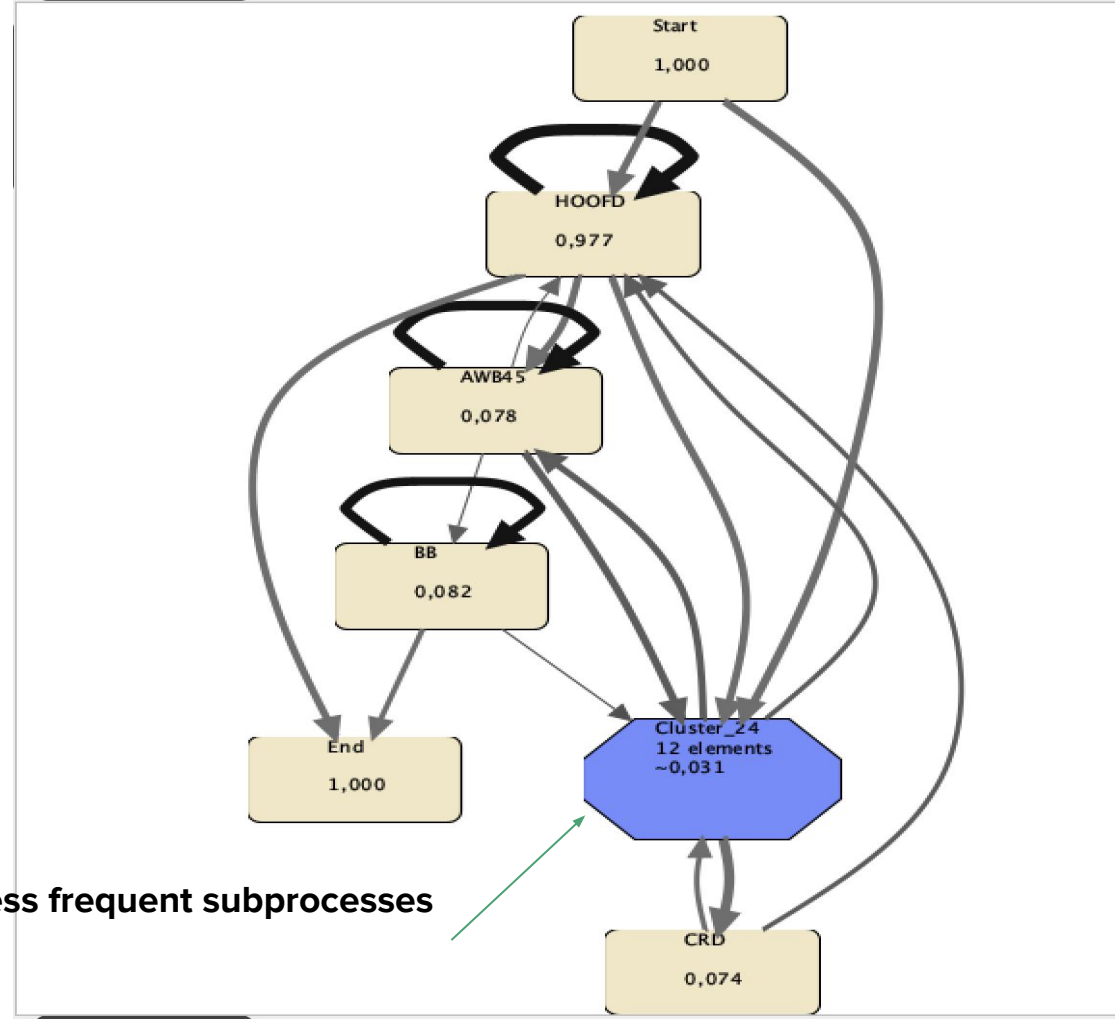
Path : 50 %



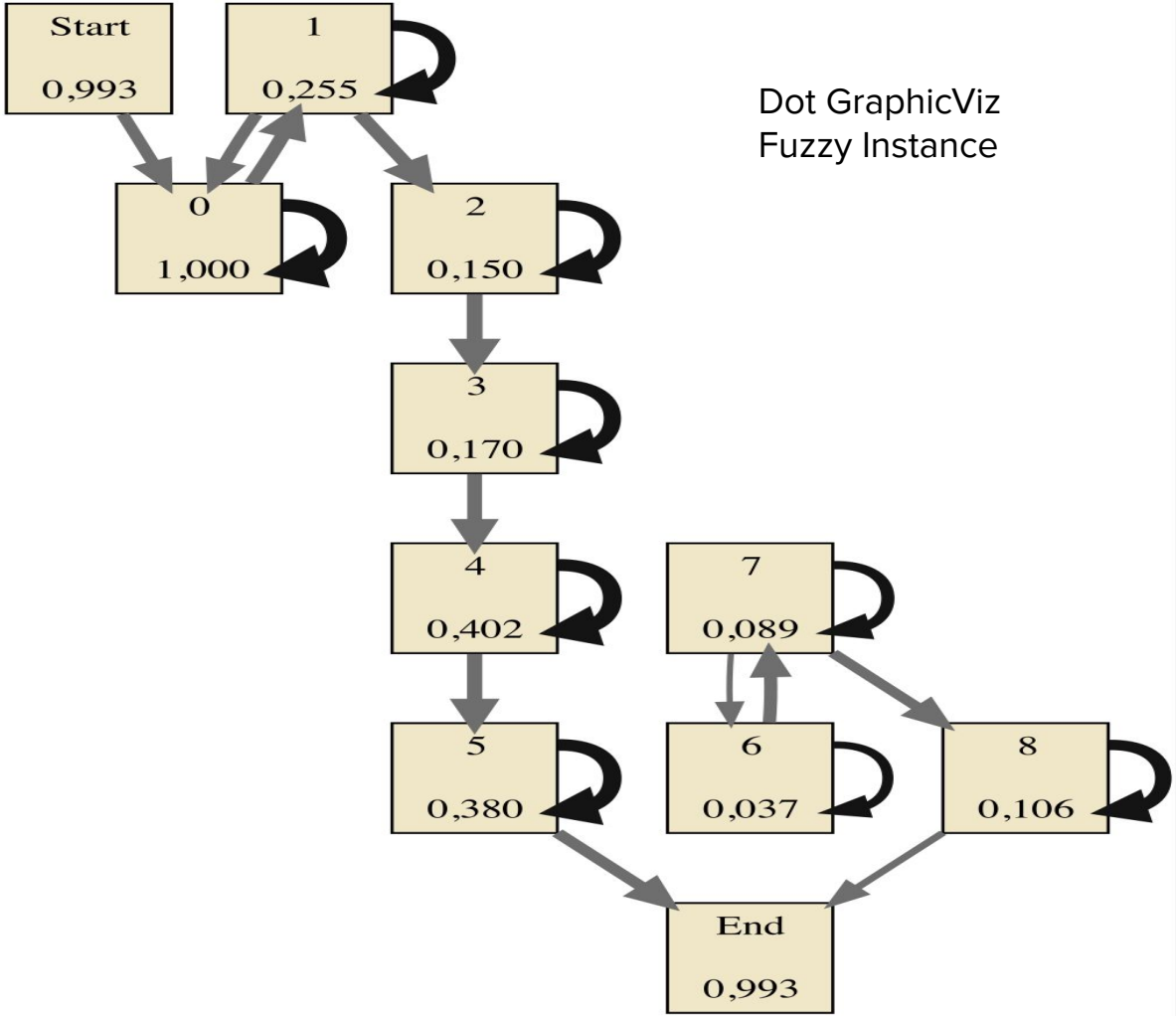
Action code level
Activity level : 50%
Path : 0%

Model 2 - Fuzzy Miner

Action Code Level



Clustered less frequent subprocesses



Dot GraphicViz
Fuzzy Instance

Phase Level Model

End