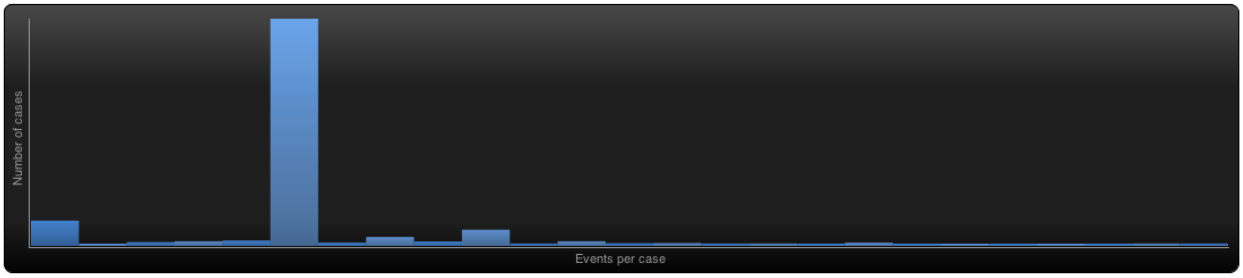
Project Title: Peer-graded Assignment (Applying Process Mining on Real Data) by jayasimha

A: Open the event log ('Receipt phase of an environmental permit application process (*WABO*) CoSeLoG project.fbt') in Disco and switch to the 'Statistics' view.

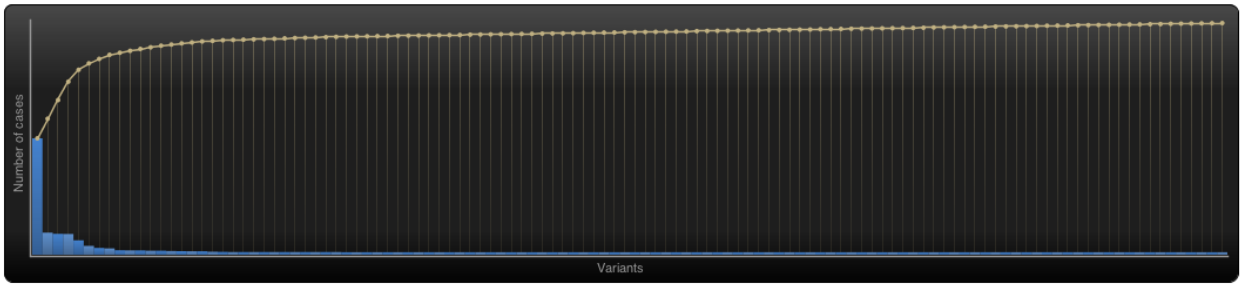
Without switching to other views, use the statistics view to answer the following three sub questions:

1. How many events are there on average per case?
2. Can you indicate whether each case seems to be unique or whether many cases follow the same activity sequence?
3. What is the main observation that can be made from the 'Events over time' graph?

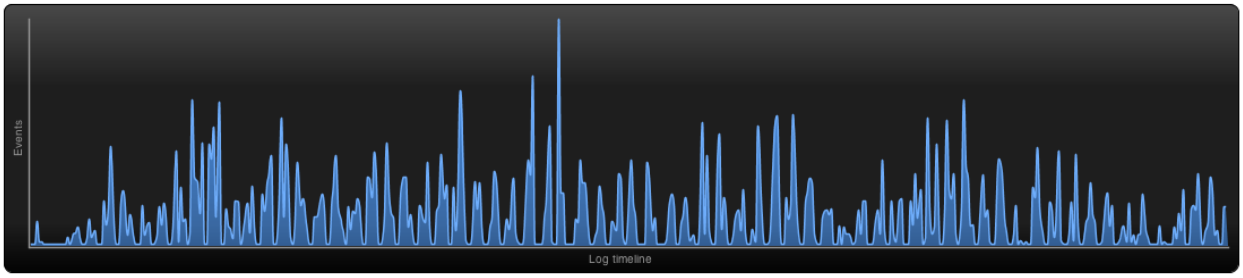
A1: In the statistics view different graphs are shown. When I see 'Events over time' button, I observed that 8577 events are found in 1434 cases, so by taking average, 8577 divided by 1434 = 5.98. So we can see that there are 6 events per case. Pic for A1:



A2: About 50% of all the cases follow single activity sequence, there are 116 total sequence variants. With a total of 1434 cases. Most of the cases are not unique, but variants 31 to 116 are single case variants. So, these cases follow a unique activity sequence. Pic for A2:



A3: If we observe, the amount of activities per day are in a range of 0 to 147. There are many groups of 2 days or more with 0 activities. Probably, the weekends and holidays. Pic for A3:



B While still in Disco, switch to the 'map' view to display a process map.

Using the map view, change the activity and path detail settings in order to create a comprehensible process map (e.g. a process map that could be printed on one A4 or letter paper or shown on a single computer screen while still being readable in full).

1. Discuss this process map, what is the main process?
2. Which activities and paths between activities are frequent?

In your answer, include the settings you used for both the activity and path sliders.

When we observe the process map pane a map for 100 percent of activities and 0 percent of path is shown. This map is not comprehensible, it contains many different activities. On the other hand only a minor number of traces finish at the end event. So the number of activities has to be decreased and the number of paths increased.

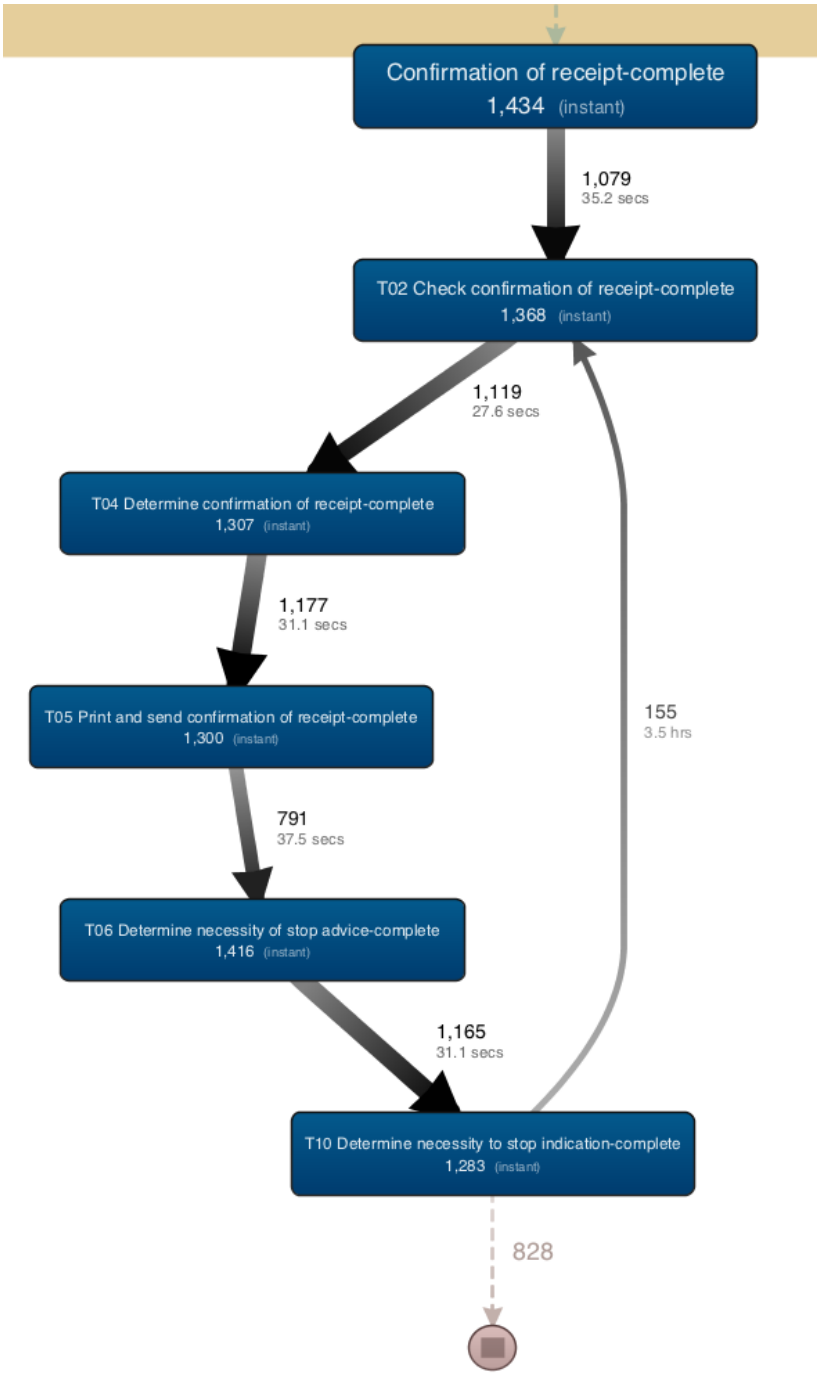
Among all, the best compromise seems to be 50% of activities and 17% of paths. So nearly 90% of traces finish at the end event and the most important activities are visible. About 90 % of all the cases contain these 6 activities:

A. major parts of the main process are 'Confirmation of receipt' and 'Determine necessity of stop'. 'Confirmation of receipt' is self-declaring, 'necessity of stop' not at all, so it is difficult to comprehend what might be stopped.

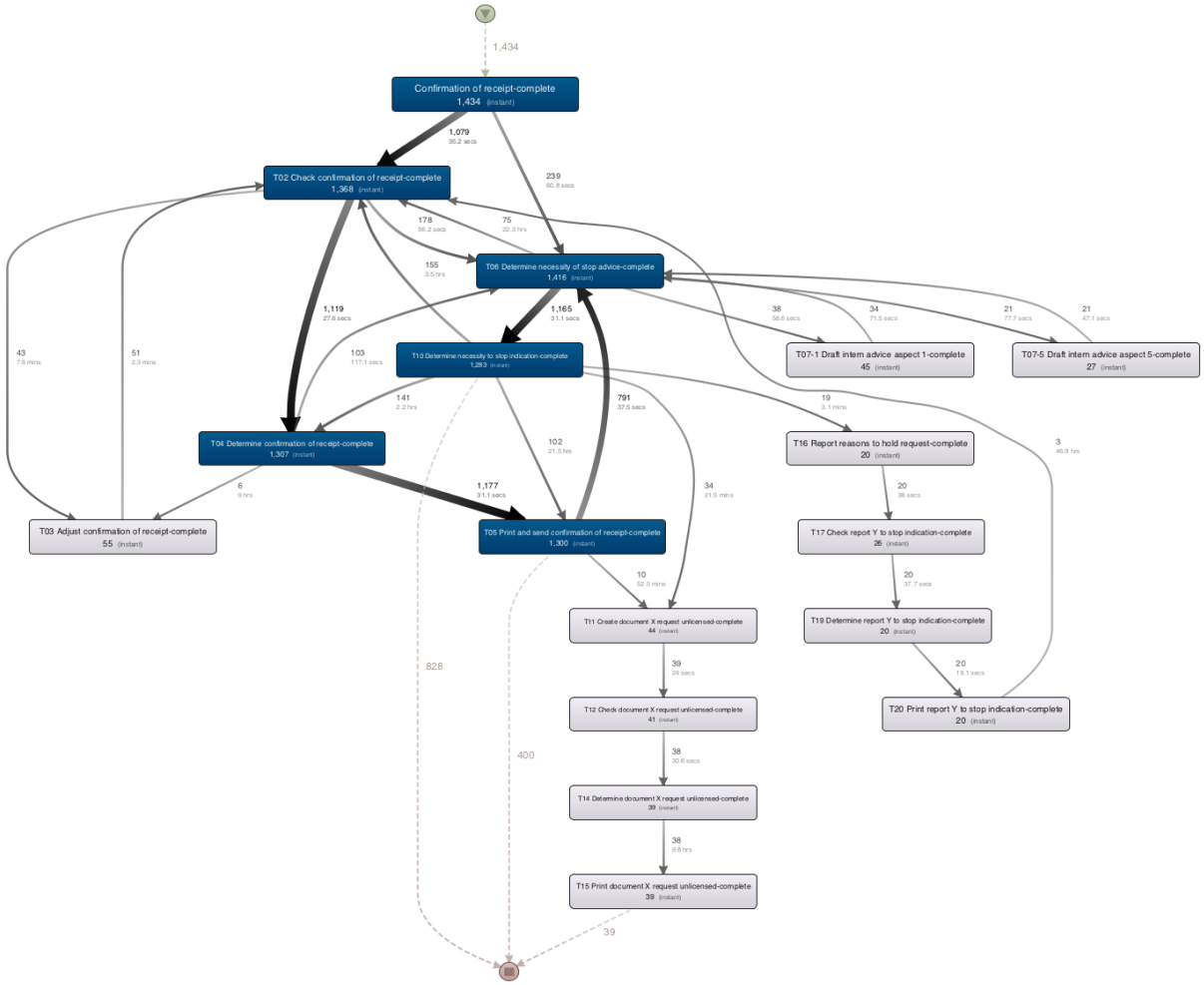
B. The main process consists of 2 groups of most frequent activities, conducted a few times parallel but most times in sequence:

1. Confirmation of receipt
2. Check confirmation of receipt
3. Determine confirmation of receipt
4. Print and sent confirmation of receipt
5. Determine necessity of stop advice
6. Determine necessity of stop indication

B1. Created in Disco.



B2.



C: While still in Disco, and while using the same process map (e.g. do not change the activity and path settings), switch to the performance projection.

1. Discuss where the process takes most time, e.g. where there are possibilities for improvement. Relate these times (of the bottlenecks) to the time spent in other parts of the process. In other words, discuss how severe the bottleneck is with respect to the time spent on other activities.
2. Also explicitly mention the performance metric chosen (e.g. total, mean, median, or max) and why you have chosen this setting.

Answer: Switching to the performance projection and keeping the sliders at 50% activities and 17% path. Setting Show to ‘Mean duration’ and Adding as Secondary ‘Absolute frequencies’. we can see:

1. Most of the time spent affecting a majority of cases is the “T06 determine necessity to stop advice activity” as can be seen in the graph, 3.1days in 791 events. If this could be shortend

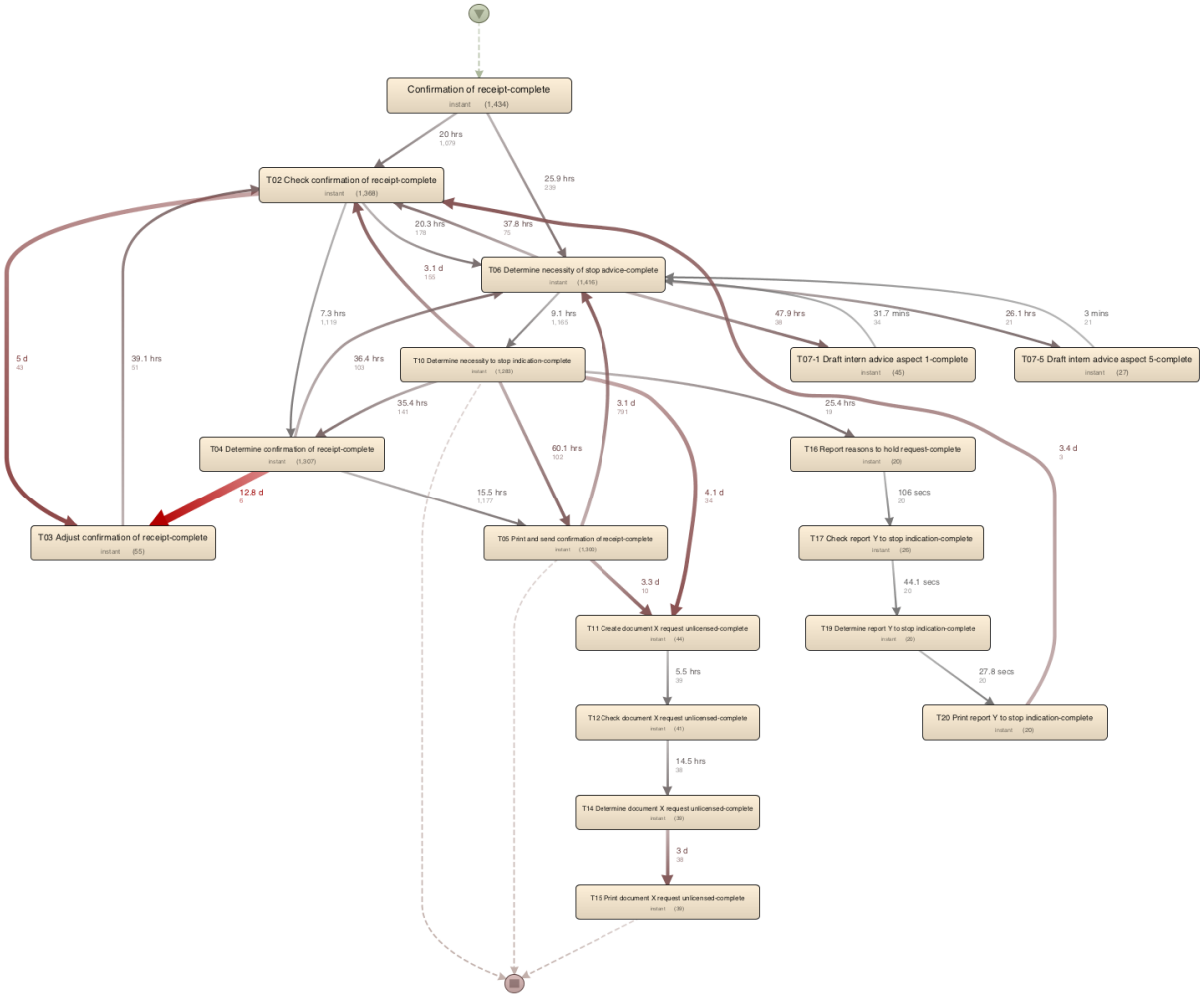
significantly, the response times would increase much and stake holders would have to wait very less. There are other even longer lasting events, but those are happening a lot less often so they are not significant as high as the above mentioned.

2. Actually, there are 2 important perspectives when we analyze process durations:

- Total durations of an activity for all cases is the relevant indicator if you want to find out where most effort went into, which relates typically which activities are the most expensive. Depending on whether these activities are evenly distributed over many parallel workers or not the total duration significantly indicates extraordinary long waiting times as well.
- Mean duration of activities and cases (full traces) is relevant if you want to find out how long typically external stake holders of a case have to wait until it is finished.

My focus here is on the external perspective - what are major factors for long case run times?

In this process mining case study total time and long mean run time are found at the same activity (79.6 mth, 791 activities). So adressing this in an improvement process change would be benefitial for both external stake holders and costs.



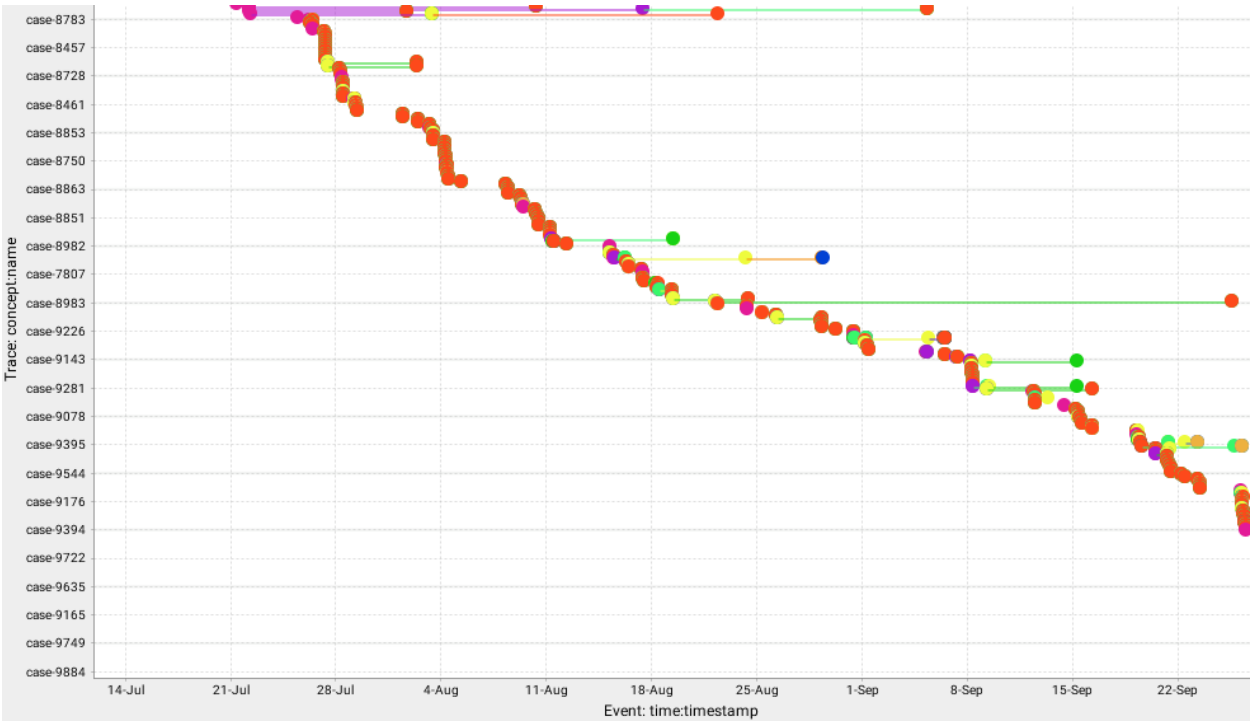
D: Now load the original event log in ProM. Visualize the event log using the Dotted Chart or XDottedChart visualizer (by pressing the 'eye'-icon with the event log selected and switching to the Dotted Chart or XDottedChart visualizer).

Using the Dotted Chart, answer the following questions:

1. Is the arrival rate of new cases constant? If not, when are there fluctuations? If yes, how can we see this from the Dotted Chart?
2. Can you observe a change in the global process?

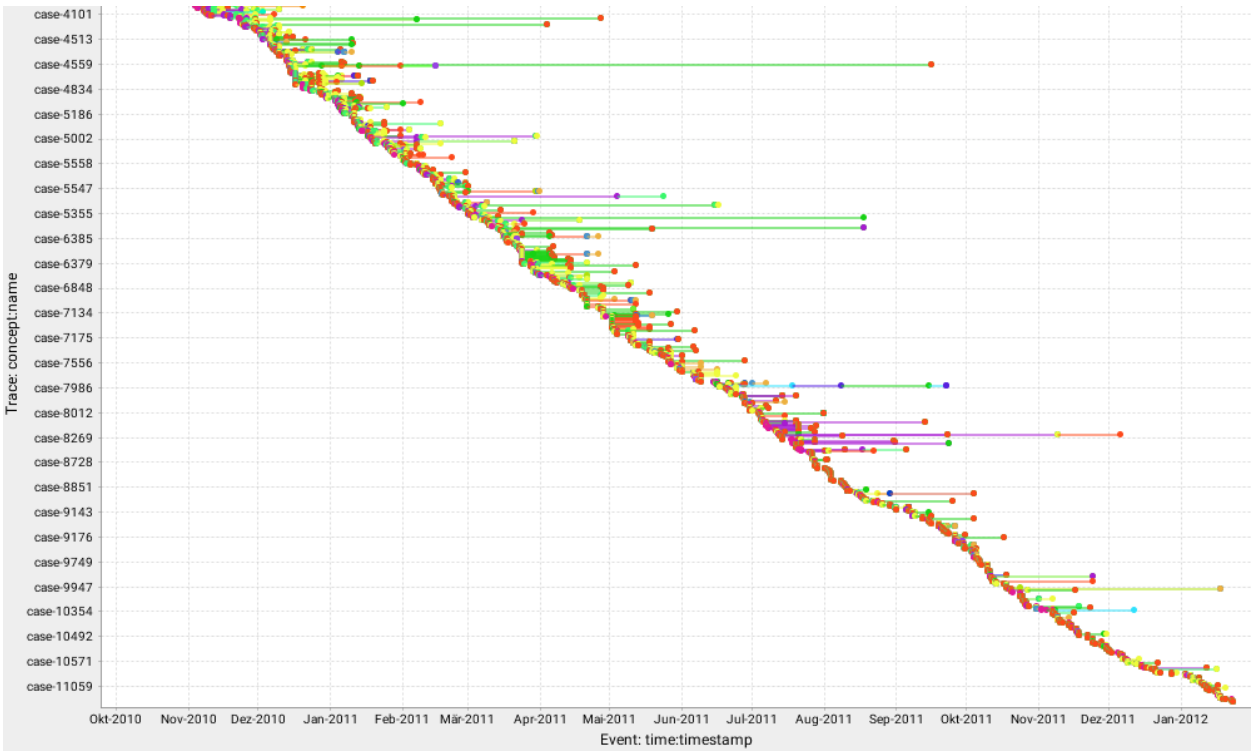
Note that you don't need to change the component, time or coloring settings. You can however re-sort the traces on the time of the first event, and zoom in or out if you want.

Loading the file Receipt phase of an environmental permit application process (*WABO*) CoSeLoGproject.xes.gz into Prom67 and visualizing it by a Dotted Chart diagram and while sorting by time: start Time of Trace gives an overview of case start rate. Magnified here:



From that diagram can be concluded:

- there are small fluctuations in the arrival rate, there are no new cases on weekends.
- there is a general trend for a slightly lower rate of new cases from around 20th of August



Looking at all traces we can see:

- there are less activities per case since about case 8800
- the total time per trace looks shorter since then too
- there are fewer very long running traces since 8800 as well this might have been caused by a change in the process handling within the organization:

E: You are now asked to discover a Petri net on the event log. However, the unfiltered event log results in an incomprehensible Petri net. Therefore, you are allowed to run the 'Filter log using simple heuristics' plug-in once on the original event log to discover a Petri net on the filtered event log.

- Clearly indicate which settings you have used for the 'Filter log using simple heuristics' plug-in.
- Explicitly motivate the filtering settings chosen, why did you pick this percentage or selection of activities?
- Discuss and argue which plug-in (or chain of plug-ins) you have used to discover a Petri net, for instance by comparing two or more plug-in results and arguing why one of the Petri nets is better.
- Explain the (best) Petri net: what is the main process and what are notable parts of the Petri net?

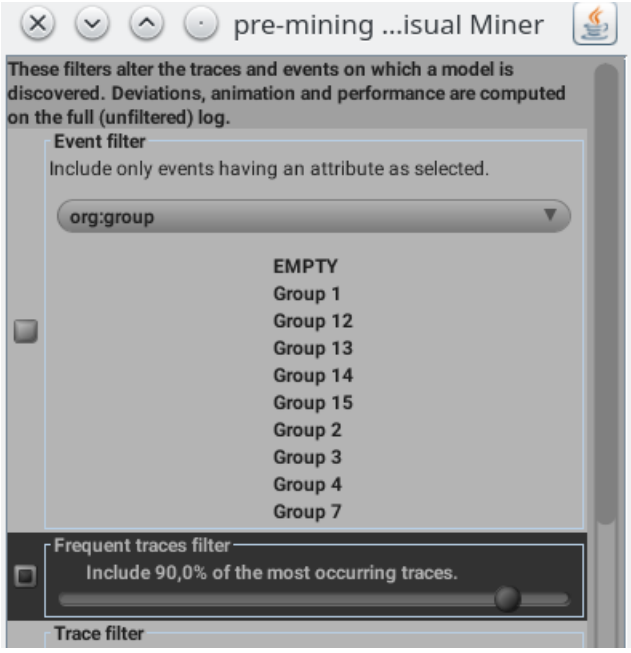
Note that this question requires you to experiment with different filtering settings and discovery plug-ins. You are not required to describe everything you have tried but found unsuccessful. Only describe the successful combination of plug-ins and its result(s) and argue why your final result is 'good'.

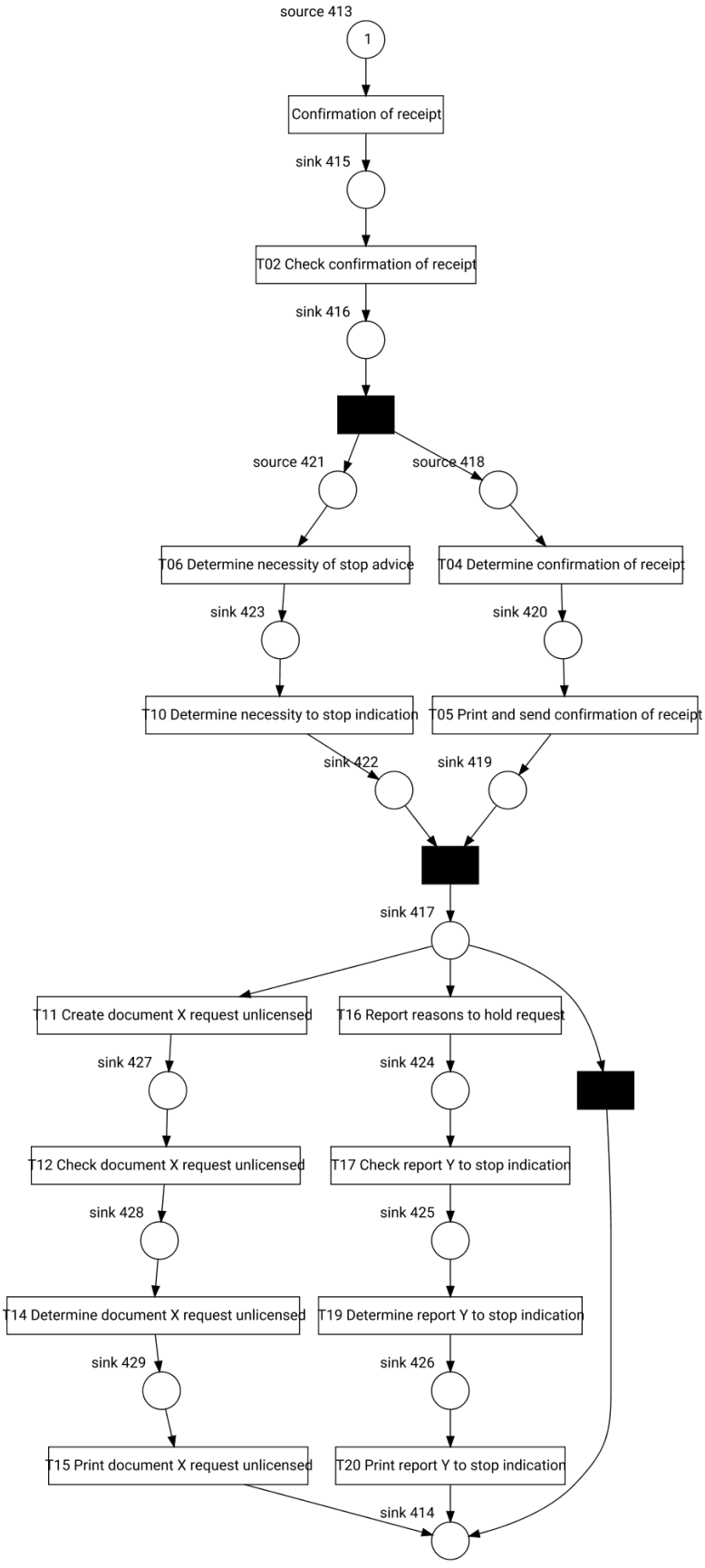
The important step to find a comprehensible Petri net from the supplied log data is to filter out classes of traces with only a few class members but which are complicating the process model without giving extra insights. The big challenge is setting the filters to such a level that an optimum of the real process variants is captured.

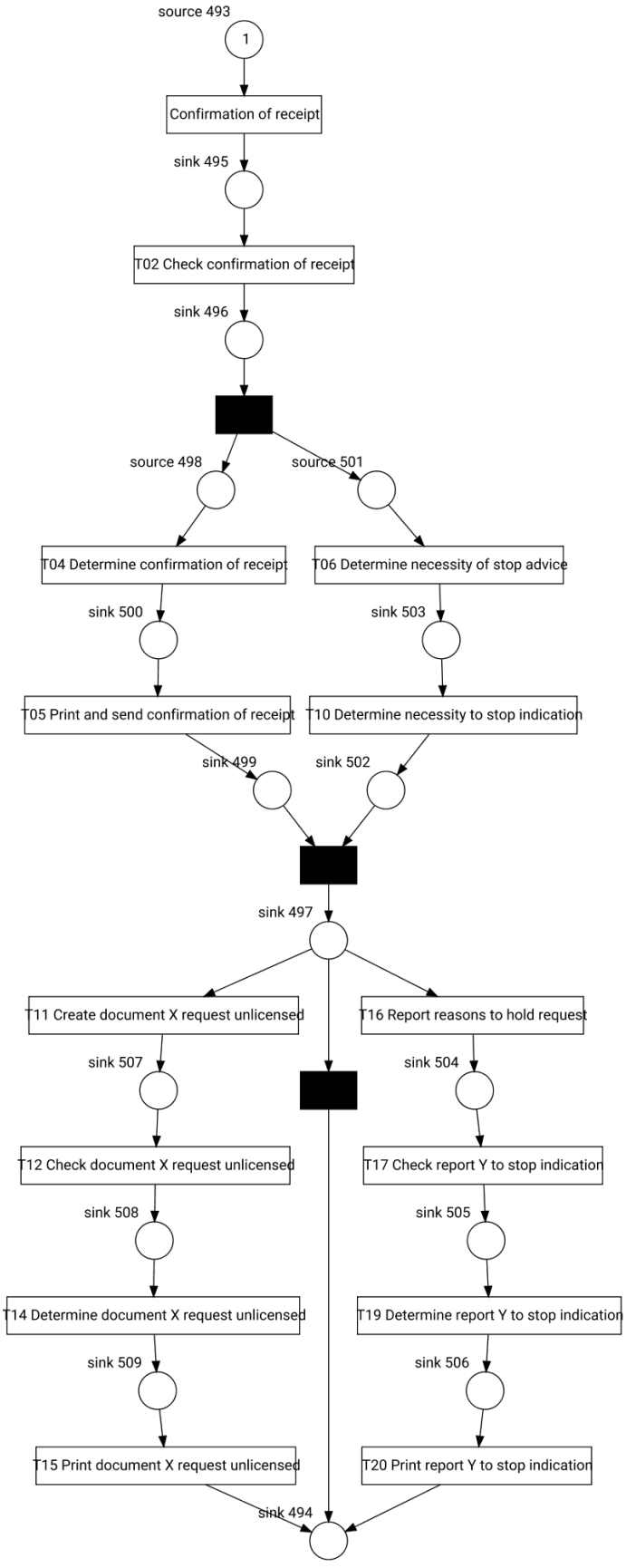
Whether a class of traces is included or not for a good process model should not depend on miner chosen for a Petri net. But there are differences in features a process discovery method is able to guarantee. Book chapter 7.5 “Inductive Mining” states “process trees are sound by construction”. So a good starting point is using the Inductive Visual Miner Plugin within ProM to discover a process tree and convert it to a Petri net thereafter.

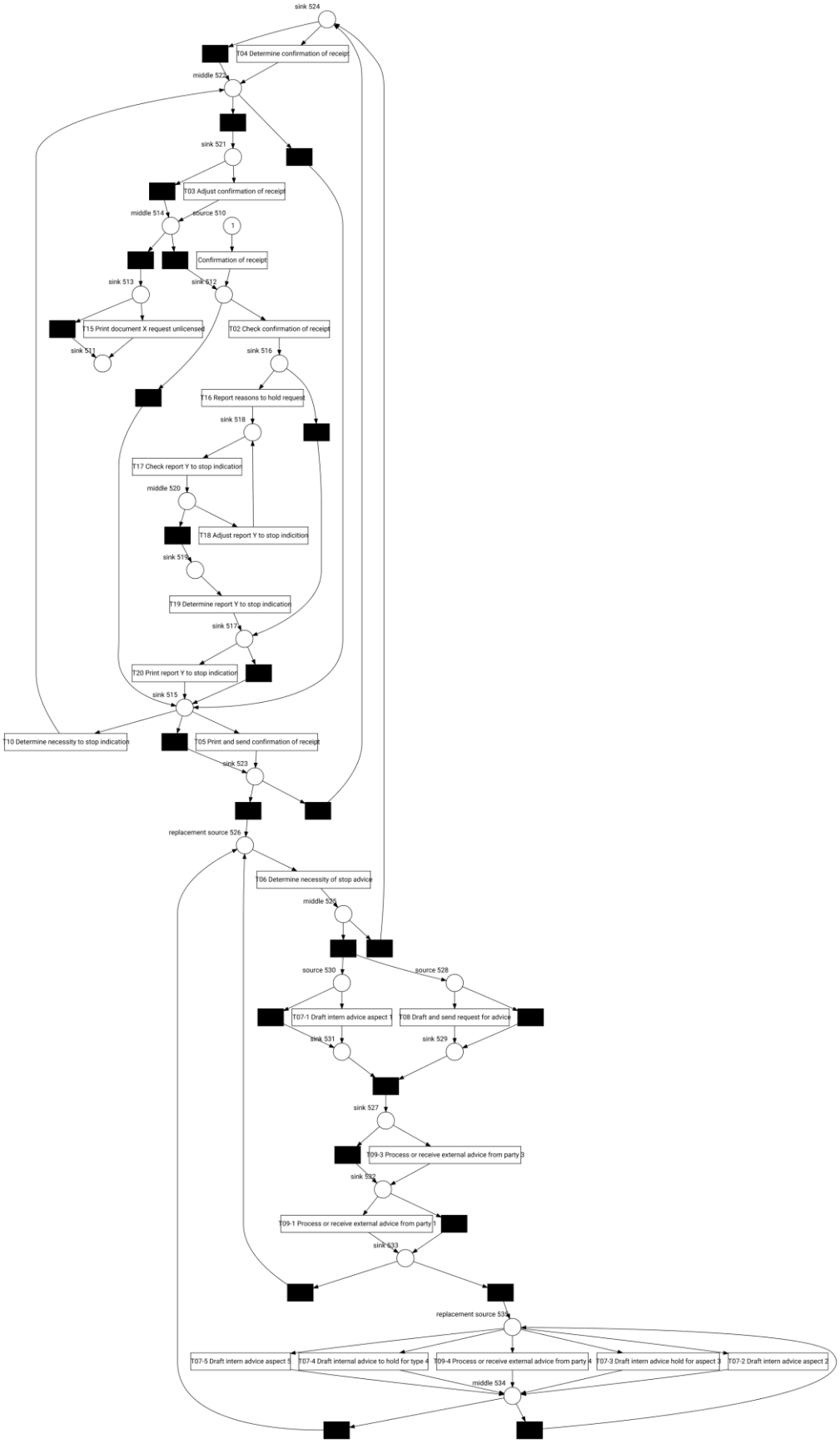
- Apply Inductive Visual Miner Plugin to the data set with default settings.
- Convert the Process tree to a Petri net
- Evaluate the Petri net
- If Petri net is satisfactory - finish
- Set pre mining filter to get a more comprehensible tree
- Repeat loop from step 2.

We can see the chart with defaults in next page:

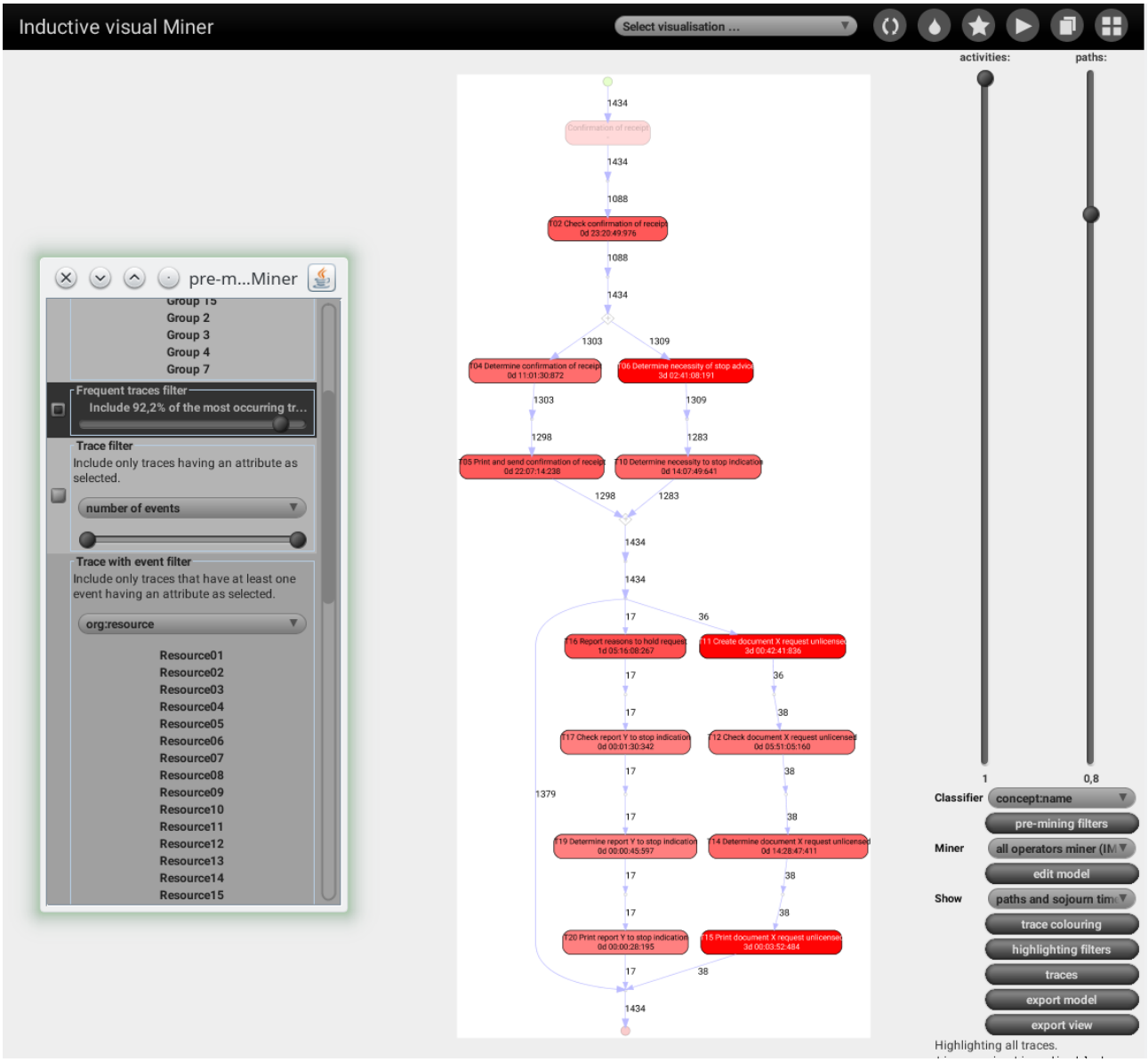


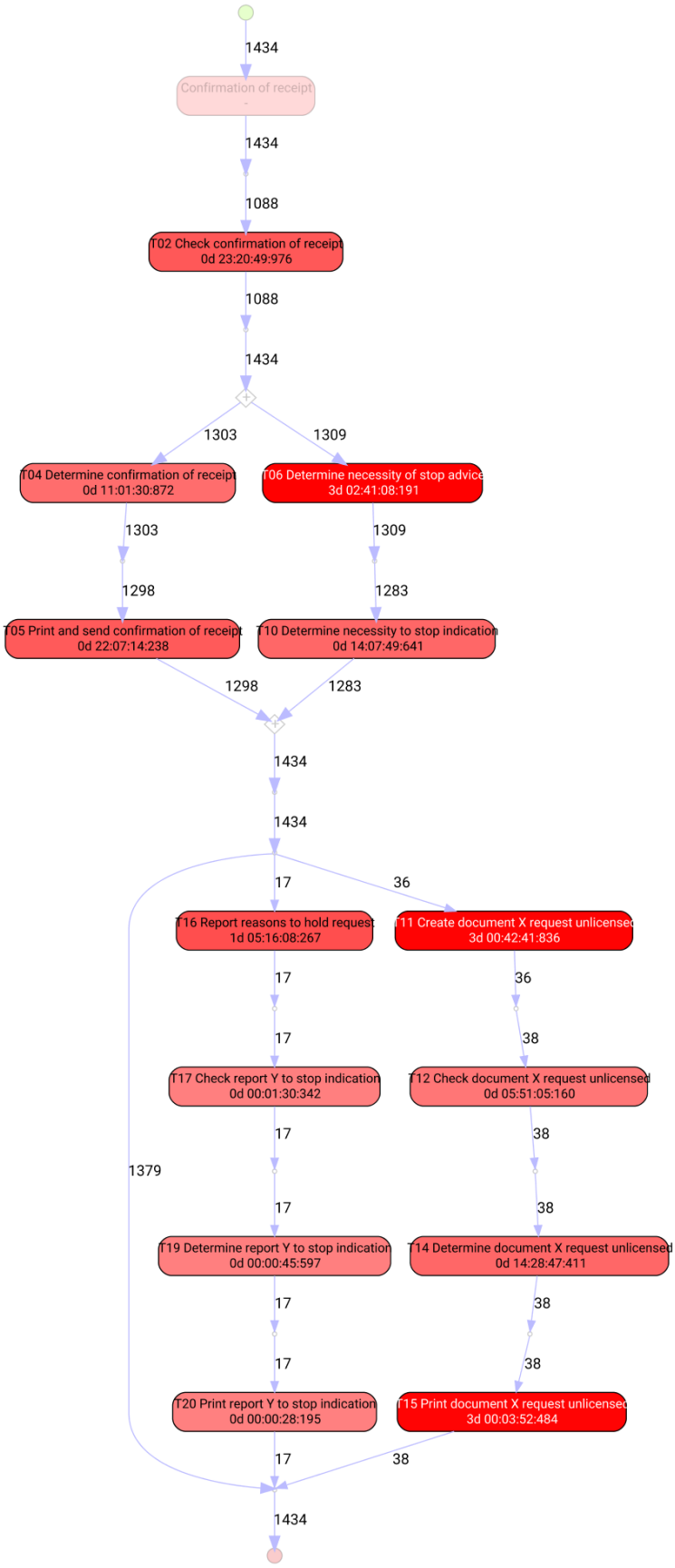

source 413
1
Confirmation of receipt
sink 415
T02 Check confirmation of receipt
sink 416
source 421
source 418
T06 Determine necessity of stop advice
T04 Determine confirmation of receipt
sink 423
sink 420
T10 Determine necessity to stop indication
T05 Print and send confirmation of receipt
sink 422
sink 419
sink 417
T11 Create document X request unlicensed
T16 Report reasons to hold request
sink 427
sink 424
T12 Check document X request unlicensed
T17 Check report Y to stop indication
sink 428
sink 425
T14 Determine document X request unlicensed
T19 Determine report Y to stop indication
sink 429
sink 426
T15 Print document X request unlicensed
T20 Print report Y to stop indication
sink 414

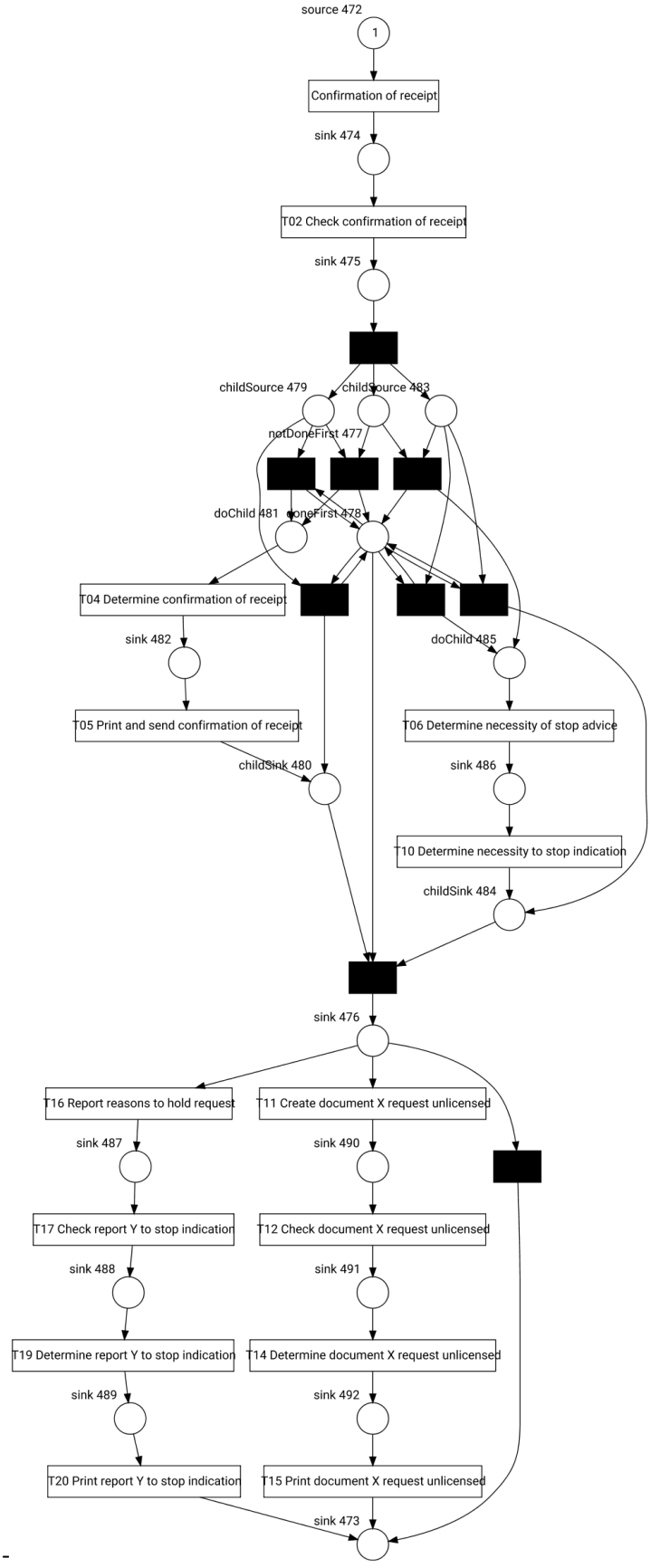

source 493
1
Confirmation of receipt
sink 495
T02 Check confirmation of receipt
sink 496
source 498
source 501
T04 Determine confirmation of receipt
T06 Determine necessity of stop advice
sink 500
sink 503
T05 Print and send confirmation of receipt
T10 Determine necessity to stop indication
sink 499
sink 502
sink 497
T11 Create document X request unlicensed
T16 Report reasons to hold request
sink 507
sink 504
T12 Check document X request unlicensed
T17 Check report Y to stop indication
sink 508
sink 505
T14 Determine document X request unlicensed
T19 Determine report Y to stop indication
sink 509
sink 506
T15 Print document X request unlicensed
T20 Print report Y to stop indication
sink 494


sink 524
T04 Determine confirmation of receipt
middle 522
sink 521
T03 Adjust confirmation of receipt
middle 514
source 510
1
Confirmation of receipt
sink 513
sink 512
T15 Print document X request unlicensed
T02 Check confirmation of receipt
sink 511
sink 516
T16 Report reasons to hold request
sink 518
T17 Check report Y to stop indication
middle 520
T18 Adjust report Y to stop indicition
sink 519
T19 Determine report Y to stop indication
sink 517
T20 Print report Y to stop indication
sink 515
T10 Determine necessity to stop indication
T05 Print and send confirmation of receipt
sink 523
replacement source 526
T06 Determine necessity of stop advice
middle 525
source 530
source 528
T07-1 Draft intern advice aspect 1
T08 Draft and send request for advice
sink 531
sink 529
sink 527
T09-3 Process or receive external advice from party 3
sink 532
T09-1 Process or receive external advice from party 1
sink 533
replacement source 535
T07-5 Draft intern advice aspect 5
T07-4 Draft internal advice to hold for type 4
T09-4 Process or receive external advice from party 4
T07-3 Draft intern advice hold for aspect 3
T07-2 Draft intern advice aspect 2
middle 534

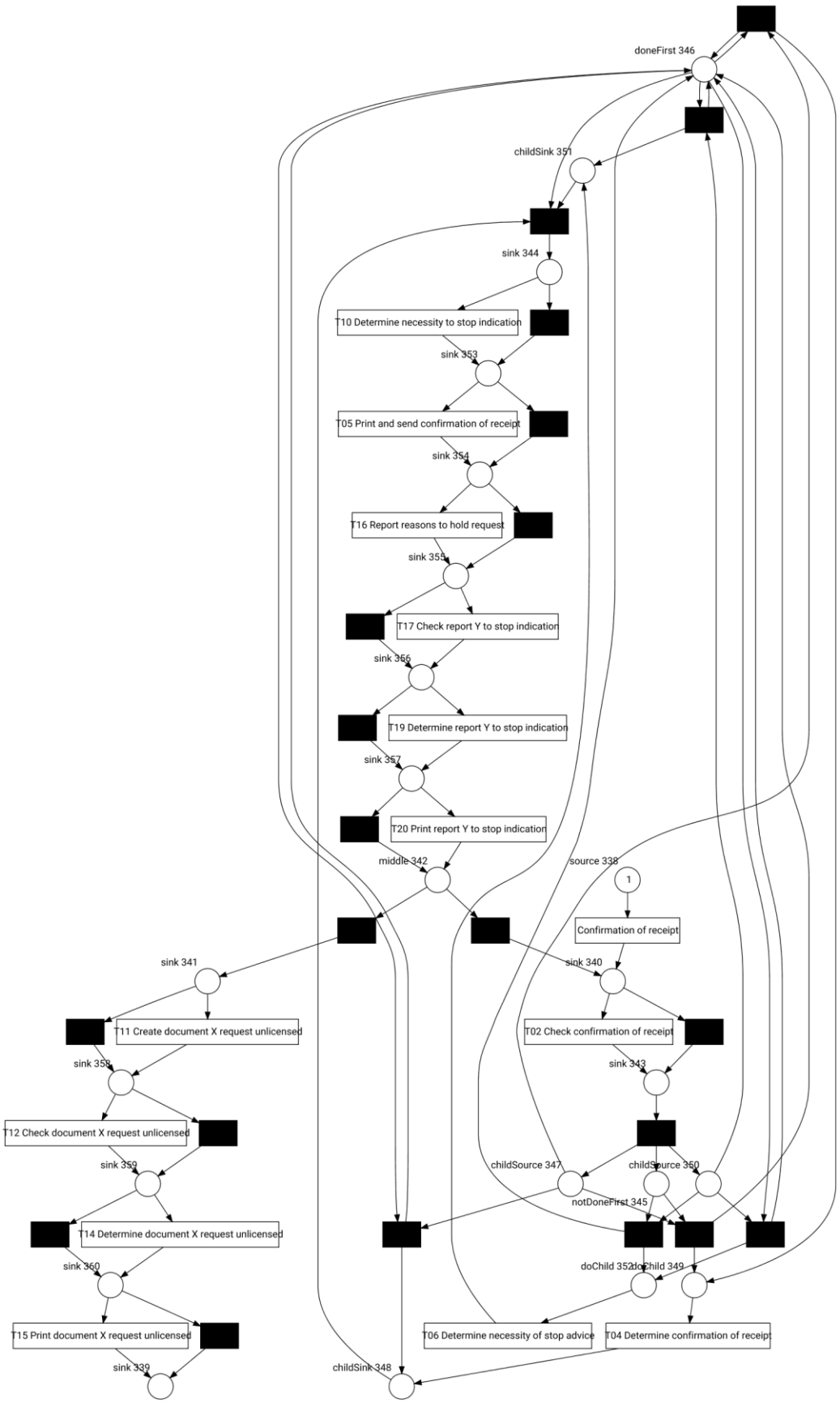
2. The Petri net is not comprehensible, it contains many hidden nodes
3. Set the trace prefilter to only include 90% of the most occurring traces. which is a result of this mined model, with parameters set up to be verified on screencopy




1434
Confirmation of receipt
1434
1088
T02 Check confirmation of receipt
0d 23:20:49:976
1088
1434
1303
1309
T04 Determine confirmation of receipt
0d 11:01:30:872
T06 Determine necessity of stop advice
3d 02:41:08:191
1303
1309
1298
1283
T05 Print and send confirmation of receipt
0d 22:07:14:238
T10 Determine necessity to stop indication
0d 14:07:49:641
1298
1283
1434
1434
17
36
T16 Report reasons to hold request
1d 05:16:08:267
T11 Create document X request unlicensed
3d 00:42:41:836
17
36
17
38
T17 Check report Y to stop indication
0d 00:01:30:342
T12 Check document X request unlicensed
0d 05:51:05:160
17
38
1379
17
38
T19 Determine report Y to stop indication
0d 00:00:45:597
T14 Determine document X request unlicensed
0d 14:28:47:411
17
38
17
38
T20 Print report Y to stop indication
0d 00:00:28:195
T15 Print document X request unlicensed
3d 00:03:52:484
17
38
1434


source 472
1
Confirmation of receipt
sink 474
T02 Check confirmation of receipt
sink 475
childSource 479
childSource 483
notDoneFirst 477
doChild 481
doneFirst 478
T04 Determine confirmation of receipt
sink 482
doChild 485
T05 Print and send confirmation of receipt
T06 Determine necessity of stop advice
childSink 480
sink 486
T10 Determine necessity to stop indication
childSink 484
sink 476
T16 Report reasons to hold request
T11 Create document X request unlicensed
sink 487
sink 490
T17 Check report Y to stop indication
T12 Check document X request unlicensed
sink 488
sink 491
T19 Determine report Y to stop indication
T14 Determine document X request unlicensed
sink 489
sink 492
T20 Print report Y to stop indication
T15 Print document X request unlicensed
sink 473

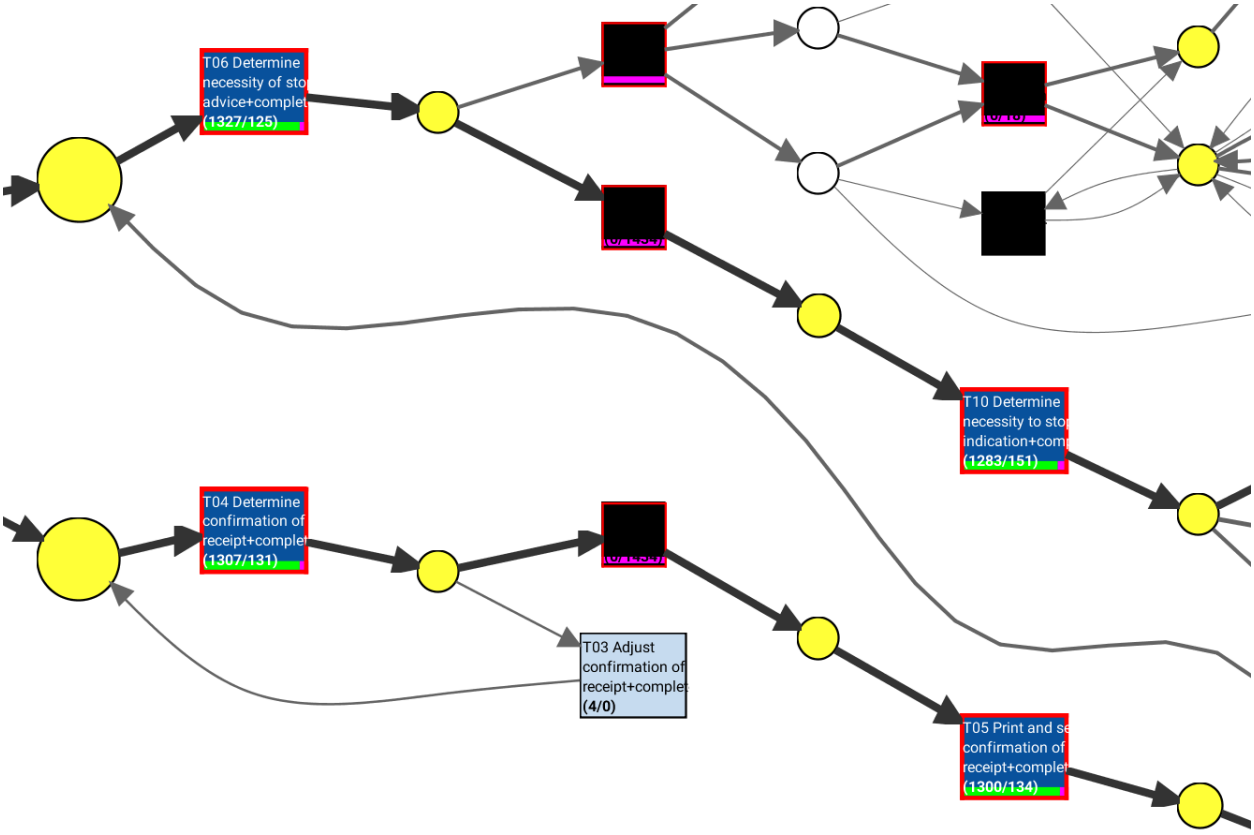
-


doneFirst 346
childSink 351
sink 344
T10 Determine necessity to stop indication
sink 353
T05 Print and send confirmation of receipt
sink 354
T16 Report reasons to hold request
sink 355
T17 Check report Y to stop indication
sink 356
T19 Determine report Y to stop indication
sink 357
T20 Print report Y to stop indication
middle 342
source 338
1
Confirmation of receipt
sink 341
sink 340
T11 Create document X request unlicensed
T02 Check confirmation of receipt
sink 358
sink 343
T12 Check document X request unlicensed
sink 359
childSource 347
childSource 350
notDoneFirst 345
T14 Determine document X request unlicensed
sink 360
doChild 352
doChild 349
T15 Print document X request unlicensed
T06 Determine necessity of stop advice
T04 Determine confirmation of receipt
sink 339
childSink 348

F: The organization has a process model that describes the 'should be' process (i.e. a normative process model). Load the file 'normativeModel.pnml' into ProM and apply conformance checking on this process model, and on the full unfiltered original event log.

Include a screenshot of the part of the normative process model, with the conformance information projected onto it, that shows where most of the deviations occur. What is the replay fitness (the 'trace fitness' statistic) of the event log on the normative process model? Select the transition 'T06 Determine necessity of stop advice+complete' (on the top left of the model) and discuss its element statistics: how many times is the transition executed correctly and how many times incorrectly? Using the element statistics of transition 'T06 Determine necessity of stop advice+complete', what can you say about the (in)correct execution of this activity?

The part of the model where the most (absolute, not relative) deviations occur:



The trace fitness is calculated as 0.8425

The overall statistics of the replay are:

Stat	Value
#Cases replayed	1.434
#Synchronous ev.class (log+model)	6.887
#Skipped ev.class	543
#Unobservable ev.class	10.287
#Inserted ev.class	1.690
#Violating synchronous ev.class	0

The screenshot also shows the T06 element statistics of 1309 traces where move log and modul occur and 125 traces with move model only.

So 8.7% of all traces show a deviation between from the model.

G The final analysis you have to perform on the original event log is a resource analysis, e.g. looking at the user behavior in the event log.

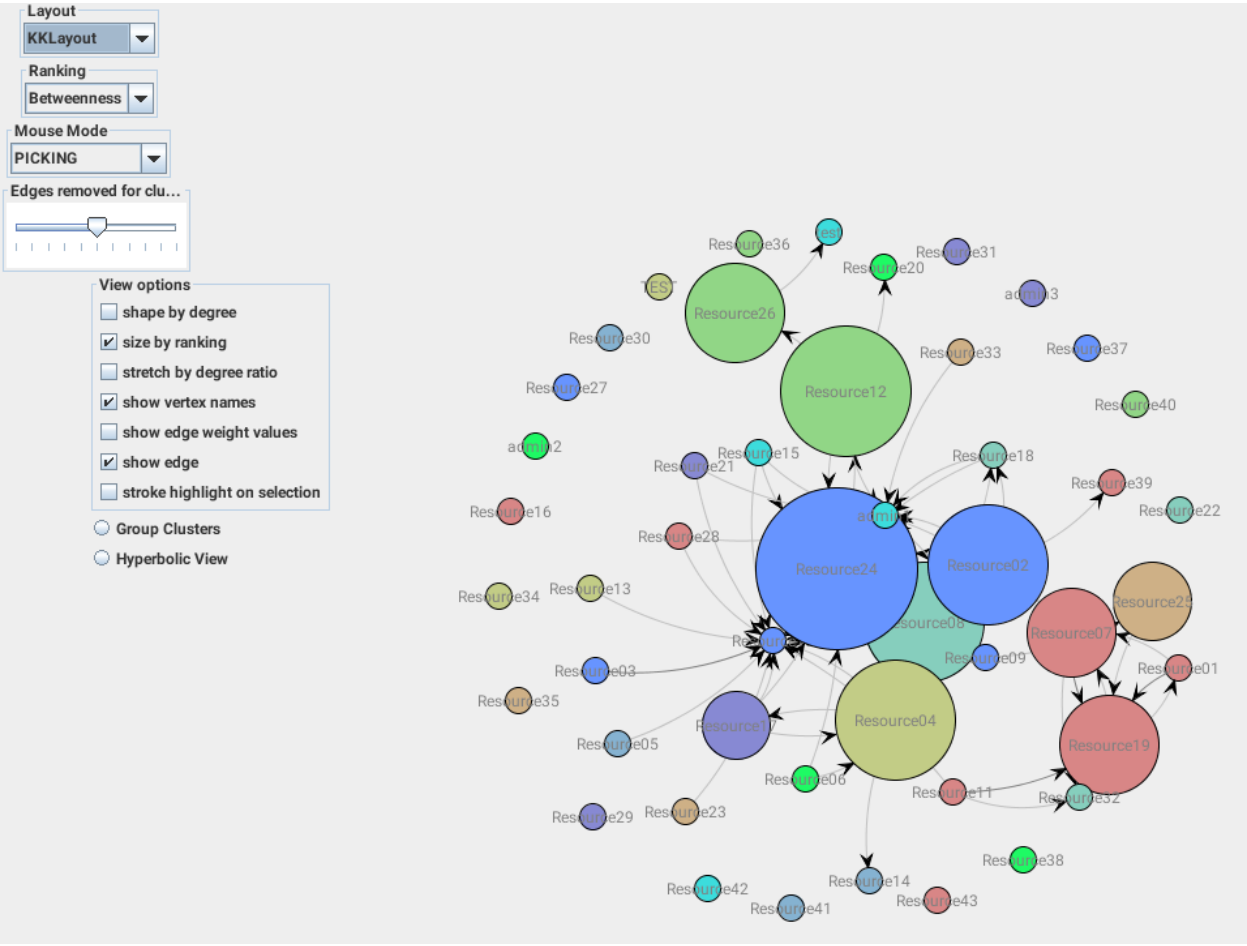
Use the plug-in 'Mine for a Subcontracting Social Network'. Note that subcontracting means that if individual *j* frequently executed an activity in-between two activities executed by individual *i*, then individual *i* subcontracted work to individual *j*.

Answer the following question using this view:

1. Can two or more groups of users be distinguished? Explicitly discuss the settings you have used in the resulting visualization.

Applied the plug-in.
Non-default settings choosen:

1. Ranking: "Betweenness" it shows ressources acting in-between
2. Edge removed for - to get distingushing colors
3. Size by ranking - to display the in-betweeness by size



We can distinguish 3 types of resource involvement:

1. With big subcontracting involvement, e.g. Resource 24, 02, 12, 08, …
2. Several with some subcontracting, e.g. 33, 15, 13, …., admin, …
3. Many with no subcontracting at all. 36, 30, 2, …

H:

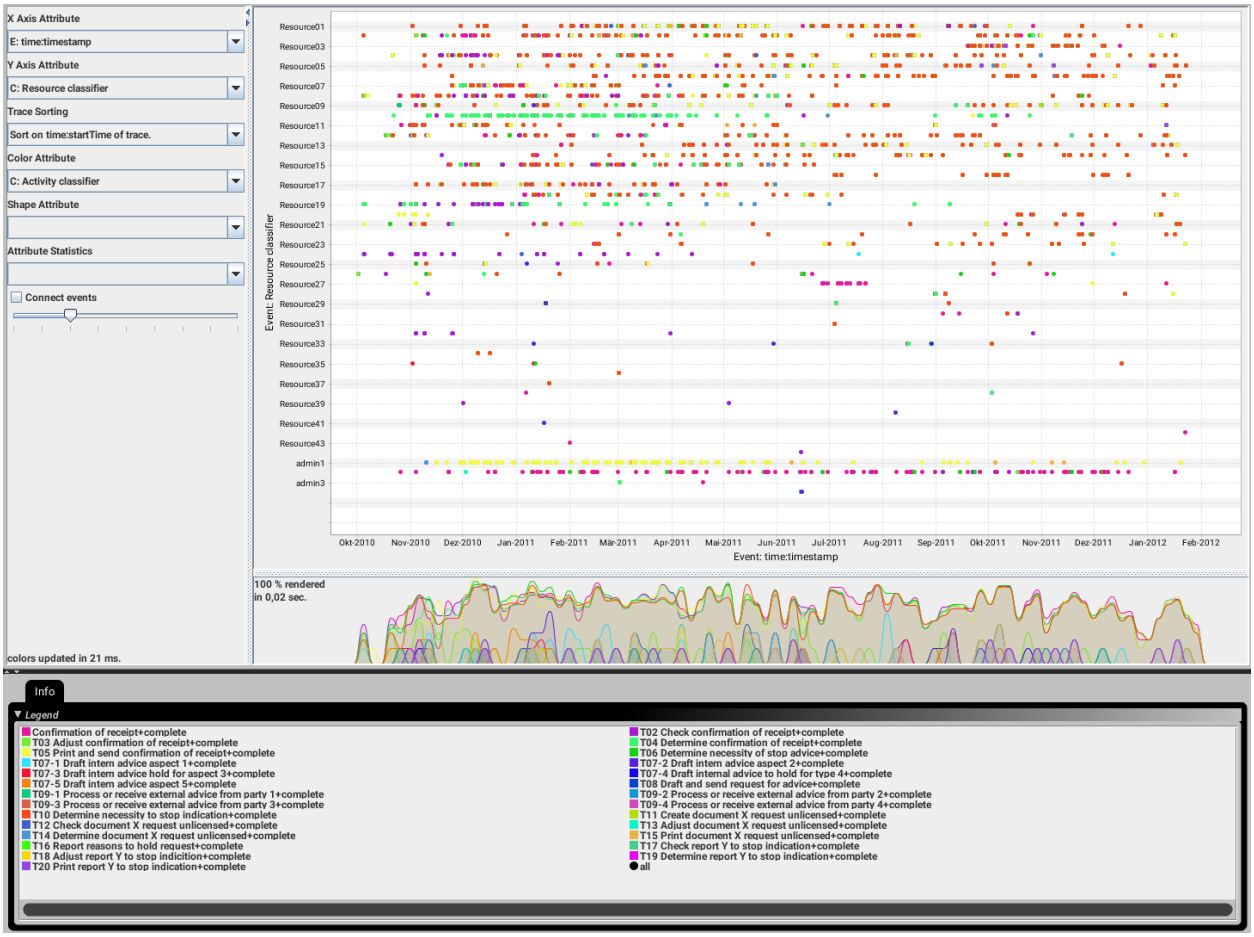
The final analysis you have to perform on the original event log is a resource analysis, e.g. looking at the user behavior in the event log.

1. Use the plug-in 'Mine for a Subcontracting Social Network'. Note that subcontracting means that if individual *j* frequently executed an activity in-between two activities executed by individual *i*, then individual *i* subcontracted work to individual *j*. Answer the following question using this view: Can two or more groups of users be distinguished? Explicitly discuss the settings you have used in the resulting visualization.
2. Again use one of the two Dotted Chart plug-ins. For the XDottedChart change the component type to 'org:resource'. If you use the Dotted Chart visualizer change the 'Y Axis Attribute' to 'C: Resource classifier' and the color attribute to 'C: Activity Classifier'. Answer the following two questions using this view:
3. Are all users executing activities from the start of the event log, or are some users joining later?
4. Are users mainly executing particular activities or are most users executing most of the activities?

Answer:

we can conclude:

1. A few resources were active only later, e.g. 03 since Sept. 2011.
2. Some resources are involved only in one or very few differnt activities. E.g. admin1 = T05.
3. About half of the resources are active most times, the other half only occasionally.
4. Some conducted only one activity at all.



J:

To conclude this assignment, briefly discuss three observations you have made during your analysis that you would like to communicate to the business user.

Think for instance of possible improvement opportunities and starting points for further investigation.

There is much effort spent in T06. Optimization of this will reduce costs and waiting times for stake holders significantly.