An introduction to Process Mining

Diogo R. Ferreira
diogo.ferreira@tecnico.ulisboa.pt
Introduction

• Example of a business process

- sequence of activities
- branching and parallel behavior
- process model vs. process instances
Introduction

• Possible questions
  • what is the most frequent path?
  • how many requests (%) get approved?
  • how much time does the process (or each activity) take?
  • does every instance comply with the model?
Introduction

• Task allocation
• More possible questions
  • what is the distribution of work, i.e. who does what?
  • how do users interact and collaborate with each other?
  • what is the performance of each user in each activity?
  • are certain resources being overloaded?
Event logs

• Example of an event log

<table>
<thead>
<tr>
<th>case id</th>
<th>task</th>
<th>user</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>u₁</td>
<td>2016-04-09 17:36:47</td>
</tr>
<tr>
<td>1</td>
<td>b</td>
<td>u₃</td>
<td>2016-04-11 09:11:13</td>
</tr>
<tr>
<td>1</td>
<td>d</td>
<td>u₆</td>
<td>2016-04-12 10:00:12</td>
</tr>
<tr>
<td>1</td>
<td>e</td>
<td>u₇</td>
<td>2016-04-12 18:21:32</td>
</tr>
<tr>
<td>1</td>
<td>f</td>
<td>u₈</td>
<td>2016-04-13 13:27:41</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>u₂</td>
<td>2016-04-14 08:56:09</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>u₃</td>
<td>2016-04-14 09:36:02</td>
</tr>
<tr>
<td>2</td>
<td>d</td>
<td>u₅</td>
<td>2016-04-15 10:16:40</td>
</tr>
<tr>
<td>1</td>
<td>g</td>
<td>u₆</td>
<td>2016-04-18 19:14:14</td>
</tr>
<tr>
<td>2</td>
<td>g</td>
<td>u₆</td>
<td>2016-04-19 15:39:15</td>
</tr>
<tr>
<td>1</td>
<td>h</td>
<td>u₂</td>
<td>2016-04-19 16:48:16</td>
</tr>
<tr>
<td>2</td>
<td>e</td>
<td>u₇</td>
<td>2016-04-20 14:39:45</td>
</tr>
<tr>
<td>2</td>
<td>f</td>
<td>u₈</td>
<td>2016-04-22 09:16:16</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>u₂</td>
<td>2016-04-25 08:39:24</td>
</tr>
<tr>
<td>2</td>
<td>h</td>
<td>u₁</td>
<td>2016-04-26 12:19:46</td>
</tr>
<tr>
<td>3</td>
<td>b</td>
<td>u₄</td>
<td>2016-04-29 10:56:14</td>
</tr>
<tr>
<td>3</td>
<td>c</td>
<td>u₁</td>
<td>2016-04-30 15:41:22</td>
</tr>
</tbody>
</table>
Event logs

• Process mining perspectives
  • control-flow perspective (*task* column)
  • organizational perspective (*user* column)
  • performance perspective (*timestamp* column)

<table>
<thead>
<tr>
<th>case id</th>
<th>task</th>
<th>user</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>u₁</td>
<td>2016-04-09 17:36:47</td>
</tr>
<tr>
<td>1</td>
<td>b</td>
<td>u₃</td>
<td>2016-04-11 09:11:13</td>
</tr>
<tr>
<td>1</td>
<td>d</td>
<td>u₆</td>
<td>2016-04-12 10:00:12</td>
</tr>
<tr>
<td>1</td>
<td>e</td>
<td>u₇</td>
<td>2016-04-12 18:21:32</td>
</tr>
<tr>
<td>1</td>
<td>f</td>
<td>u₈</td>
<td>2016-04-13 13:27:41</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>u₂</td>
<td>2016-04-14 08:56:09</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>u₃</td>
<td>2016-04-14 09:36:02</td>
</tr>
<tr>
<td>2</td>
<td>d</td>
<td>u₅</td>
<td>2016-04-15 10:16:40</td>
</tr>
<tr>
<td>1</td>
<td>g</td>
<td>u₆</td>
<td>2016-04-18 19:14:14</td>
</tr>
<tr>
<td>2</td>
<td>g</td>
<td>u₆</td>
<td>2016-04-19 15:39:15</td>
</tr>
<tr>
<td>1</td>
<td>h</td>
<td>u₂</td>
<td>2016-04-19 16:48:16</td>
</tr>
<tr>
<td>2</td>
<td>e</td>
<td>u₇</td>
<td>2016-04-20 14:39:45</td>
</tr>
<tr>
<td>2</td>
<td>f</td>
<td>u₈</td>
<td>2016-04-22 09:16:16</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>u₂</td>
<td>2016-04-25 08:39:24</td>
</tr>
<tr>
<td>2</td>
<td>h</td>
<td>u₁</td>
<td>2016-04-26 12:19:46</td>
</tr>
<tr>
<td>3</td>
<td>b</td>
<td>u₄</td>
<td>2016-04-29 10:56:14</td>
</tr>
<tr>
<td>3</td>
<td>c</td>
<td>u₁</td>
<td>2016-04-30 15:41:22</td>
</tr>
</tbody>
</table>
## Control-flow perspective

- Task transitions within each case id

<table>
<thead>
<tr>
<th>case id</th>
<th>task</th>
<th>user</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>u₁</td>
<td>2016-04-09 17:36:47</td>
</tr>
<tr>
<td>1</td>
<td>b</td>
<td>u₃</td>
<td>2016-04-11 09:11:13</td>
</tr>
<tr>
<td>1</td>
<td>d</td>
<td>u₆</td>
<td>2016-04-12 10:00:12</td>
</tr>
<tr>
<td>1</td>
<td>e</td>
<td>u₇</td>
<td>2016-04-12 18:21:32</td>
</tr>
<tr>
<td>1</td>
<td>f</td>
<td>u₈</td>
<td>2016-04-13 13:27:41</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>u₂</td>
<td>2016-04-14 08:56:09</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>u₃</td>
<td>2016-04-14 09:36:02</td>
</tr>
<tr>
<td>2</td>
<td>d</td>
<td>u₅</td>
<td>2016-04-15 10:16:40</td>
</tr>
<tr>
<td>1</td>
<td>g</td>
<td>u₆</td>
<td>2016-04-18 19:14:14</td>
</tr>
<tr>
<td>2</td>
<td>g</td>
<td>u₆</td>
<td>2016-04-19 15:39:15</td>
</tr>
<tr>
<td>1</td>
<td>h</td>
<td>u₂</td>
<td>2016-04-19 16:48:16</td>
</tr>
<tr>
<td>2</td>
<td>e</td>
<td>u₇</td>
<td>2016-04-20 14:39:45</td>
</tr>
<tr>
<td>2</td>
<td>f</td>
<td>u₈</td>
<td>2016-04-22 09:16:16</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>u₂</td>
<td>2016-04-25 08:39:24</td>
</tr>
<tr>
<td>2</td>
<td>h</td>
<td>u₁</td>
<td>2016-04-26 12:19:46</td>
</tr>
<tr>
<td>3</td>
<td>b</td>
<td>u₄</td>
<td>2016-04-29 10:56:14</td>
</tr>
<tr>
<td>3</td>
<td>c</td>
<td>u₁</td>
<td>2016-04-30 15:41:22</td>
</tr>
</tbody>
</table>
Control-flow perspective

- Transition matrix

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>g</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Control-flow perspective

• Transition graph
Control-flow perspective

• Transition graph for a larger event log

• percentage of approvals: \( \frac{5683}{1866 + 5683} \approx 75\% \)
Control-flow perspective

• Improving the visualization
  • edge thickness

  ![Diagram](image1)

• activity counts

  ![Diagram](image2)

• node coloring

  ![Diagram](image3)
Organizational perspective

• Handover of work

<table>
<thead>
<tr>
<th>case id</th>
<th>task</th>
<th>user</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>u1</td>
<td>2016-04-09 17:36:47</td>
</tr>
<tr>
<td>1</td>
<td>b</td>
<td>u3</td>
<td>2016-04-11 09:11:13</td>
</tr>
<tr>
<td>1</td>
<td>d</td>
<td>u6</td>
<td>2016-04-12 10:00:12</td>
</tr>
<tr>
<td>1</td>
<td>e</td>
<td>u7</td>
<td>2016-04-12 18:21:32</td>
</tr>
<tr>
<td>1</td>
<td>f</td>
<td>u8</td>
<td>2016-04-13 13:27:41</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>u2</td>
<td>2016-04-14 08:56:09</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>u3</td>
<td>2016-04-14 09:36:02</td>
</tr>
<tr>
<td>2</td>
<td>d</td>
<td>u5</td>
<td>2016-04-15 10:16:40</td>
</tr>
<tr>
<td>1</td>
<td>g</td>
<td>u6</td>
<td>2016-04-18 19:14:14</td>
</tr>
<tr>
<td>2</td>
<td>g</td>
<td>u6</td>
<td>2016-04-19 15:39:15</td>
</tr>
<tr>
<td>1</td>
<td>h</td>
<td>u2</td>
<td>2016-04-19 16:48:16</td>
</tr>
<tr>
<td>2</td>
<td>e</td>
<td>u7</td>
<td>2016-04-20 14:39:45</td>
</tr>
<tr>
<td>2</td>
<td>f</td>
<td>u8</td>
<td>2016-04-22 09:16:16</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>u2</td>
<td>2016-04-25 08:39:24</td>
</tr>
<tr>
<td>2</td>
<td>h</td>
<td>u1</td>
<td>2016-04-26 12:19:46</td>
</tr>
<tr>
<td>3</td>
<td>b</td>
<td>u4</td>
<td>2016-04-29 10:56:14</td>
</tr>
<tr>
<td>3</td>
<td>c</td>
<td>u1</td>
<td>2016-04-30 15:41:22</td>
</tr>
</tbody>
</table>
### Organizational perspective

- Transition matrix

<table>
<thead>
<tr>
<th></th>
<th>$u_1$</th>
<th>$u_2$</th>
<th>$u_3$</th>
<th>$u_4$</th>
<th>$u_5$</th>
<th>$u_6$</th>
<th>$u_7$</th>
<th>$u_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$u_2$</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$u_3$</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$u_4$</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$u_5$</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$u_6$</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$u_7$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>$u_8$</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Organizational perspective

• Transition graph
Organizational perspective

- Transition graph for a larger event log
### Organizational perspective

- **Working together**

<table>
<thead>
<tr>
<th>case id</th>
<th>task</th>
<th>user</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>( u_1 )</td>
<td>2016-04-09 17:36:47</td>
</tr>
<tr>
<td>1</td>
<td>b</td>
<td>( u_3 )</td>
<td>2016-04-11 09:11:13</td>
</tr>
<tr>
<td>1</td>
<td>d</td>
<td>( u_6 )</td>
<td>2016-04-12 10:00:12</td>
</tr>
<tr>
<td>1</td>
<td>e</td>
<td>( u_7 )</td>
<td>2016-04-12 18:21:32</td>
</tr>
<tr>
<td>1</td>
<td>f</td>
<td>( u_8 )</td>
<td>2016-04-13 13:27:41</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>( u_2 )</td>
<td>2016-04-14 08:56:09</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>( u_3 )</td>
<td>2016-04-14 09:36:02</td>
</tr>
<tr>
<td>2</td>
<td>d</td>
<td>( u_5 )</td>
<td>2016-04-15 10:16:40</td>
</tr>
<tr>
<td>1</td>
<td>g</td>
<td>( u_6 )</td>
<td>2016-04-18 19:14:14</td>
</tr>
<tr>
<td>2</td>
<td>g</td>
<td>( u_6 )</td>
<td>2016-04-19 15:39:15</td>
</tr>
<tr>
<td>1</td>
<td>h</td>
<td>( u_2 )</td>
<td>2016-04-19 16:48:16</td>
</tr>
<tr>
<td>2</td>
<td>e</td>
<td>( u_7 )</td>
<td>2016-04-20 14:39:45</td>
</tr>
<tr>
<td>2</td>
<td>f</td>
<td>( u_8 )</td>
<td>2016-04-22 09:16:16</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>( u_2 )</td>
<td>2016-04-25 08:39:24</td>
</tr>
<tr>
<td>2</td>
<td>h</td>
<td>( u_1 )</td>
<td>2016-04-26 12:19:46</td>
</tr>
<tr>
<td>3</td>
<td>b</td>
<td>( u_4 )</td>
<td>2016-04-29 10:56:14</td>
</tr>
<tr>
<td>3</td>
<td>c</td>
<td>( u_1 )</td>
<td>2016-04-30 15:41:22</td>
</tr>
</tbody>
</table>
Organizational perspective

- Number of cases where each pair of users have worked together

<table>
<thead>
<tr>
<th></th>
<th>$u_1$</th>
<th>$u_2$</th>
<th>$u_3$</th>
<th>$u_4$</th>
<th>$u_5$</th>
<th>$u_6$</th>
<th>$u_7$</th>
<th>$u_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$u_2$</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$u_3$</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$u_4$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$u_5$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$u_6$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$u_7$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$u_8$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Organizational perspective

- Number of cases where each pair of users have worked together
Organizational perspective

- Number of cases where each pair of users have worked together
### Organizational perspective

- Distribution of work

<table>
<thead>
<tr>
<th>case id</th>
<th>task</th>
<th>user</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>u₁</td>
<td>2016-04-09 17:36:47</td>
</tr>
<tr>
<td>1</td>
<td>b</td>
<td>u₃</td>
<td>2016-04-11 09:11:13</td>
</tr>
<tr>
<td>1</td>
<td>d</td>
<td>u₆</td>
<td>2016-04-12 10:00:12</td>
</tr>
<tr>
<td>1</td>
<td>e</td>
<td>u₇</td>
<td>2016-04-12 18:21:32</td>
</tr>
<tr>
<td>1</td>
<td>f</td>
<td>u₈</td>
<td>2016-04-13 13:27:41</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>u₂</td>
<td>2016-04-14 08:56:09</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>u₃</td>
<td>2016-04-14 09:36:02</td>
</tr>
<tr>
<td>2</td>
<td>d</td>
<td>u₅</td>
<td>2016-04-15 10:16:40</td>
</tr>
<tr>
<td>1</td>
<td>g</td>
<td>u₆</td>
<td>2016-04-18 19:14:14</td>
</tr>
<tr>
<td>2</td>
<td>g</td>
<td>u₆</td>
<td>2016-04-19 15:39:15</td>
</tr>
<tr>
<td>1</td>
<td>h</td>
<td>u₂</td>
<td>2016-04-19 16:48:16</td>
</tr>
<tr>
<td>2</td>
<td>e</td>
<td>u₇</td>
<td>2016-04-20 14:39:45</td>
</tr>
<tr>
<td>2</td>
<td>f</td>
<td>u₈</td>
<td>2016-04-22 09:16:16</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>u₂</td>
<td>2016-04-25 08:39:24</td>
</tr>
<tr>
<td>2</td>
<td>h</td>
<td>u₁</td>
<td>2016-04-26 12:19:46</td>
</tr>
<tr>
<td>3</td>
<td>b</td>
<td>u₄</td>
<td>2016-04-29 10:56:14</td>
</tr>
<tr>
<td>3</td>
<td>c</td>
<td>u₁</td>
<td>2016-04-30 15:41:22</td>
</tr>
</tbody>
</table>
Organizational perspective

• Distribution of work
### Performance perspective

- **Timestamp difference between events**

<table>
<thead>
<tr>
<th>case id</th>
<th>task</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>2016-04-09 17:36:47</td>
</tr>
<tr>
<td>1</td>
<td>b</td>
<td>2016-04-11 09:11:13</td>
</tr>
<tr>
<td>1</td>
<td>d</td>
<td>2016-04-12 10:00:12</td>
</tr>
<tr>
<td>1</td>
<td>e</td>
<td>2016-04-12 18:21:32</td>
</tr>
<tr>
<td>1</td>
<td>f</td>
<td>2016-04-13 13:27:41</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>2016-04-14 08:56:09</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>2016-04-14 09:36:02</td>
</tr>
<tr>
<td>2</td>
<td>d</td>
<td>2016-04-15 10:16:40</td>
</tr>
<tr>
<td>1</td>
<td>g</td>
<td>2016-04-18 19:14:14</td>
</tr>
<tr>
<td>2</td>
<td>g</td>
<td>2016-04-19 15:39:15</td>
</tr>
<tr>
<td>1</td>
<td>h</td>
<td>2016-04-19 16:48:16</td>
</tr>
<tr>
<td>2</td>
<td>e</td>
<td>2016-04-20 14:39:45</td>
</tr>
<tr>
<td>2</td>
<td>f</td>
<td>2016-04-22 09:16:16</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>2016-04-25 08:39:24</td>
</tr>
<tr>
<td>2</td>
<td>h</td>
<td>2016-04-26 12:19:46</td>
</tr>
<tr>
<td>3</td>
<td>b</td>
<td>2016-04-29 10:56:14</td>
</tr>
<tr>
<td>3</td>
<td>c</td>
<td>2016-04-30 15:41:22</td>
</tr>
</tbody>
</table>
Performance perspective

- Timestamp difference between events
Performance perspective

• Timeline of events
Performance perspective

- Relative time
Performance perspective

• Activity duration

![Graphs showing activity duration over days](image)
Conclusion

• Where to go from here
  • process mining books
    • *Process Mining: Data Science in Action*, W. van der Aalst (Springer, 2016)
  • process mining courses
    • *Process Mining: Data Science in Action*  https://www.coursera.org/learn/process-mining
    • *Introduction to Process Mining with ProM*  https://www.futurelearn.com/courses/process-mining
  • process mining website
    • http://www.processmining.org
  • process mining tools: ProM, Disco, etc.