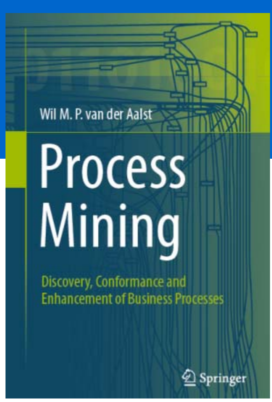
Process Mining: Data Science in Action

Learning Dependency Graphs

prof.dr.ir. Wil van der Aalst

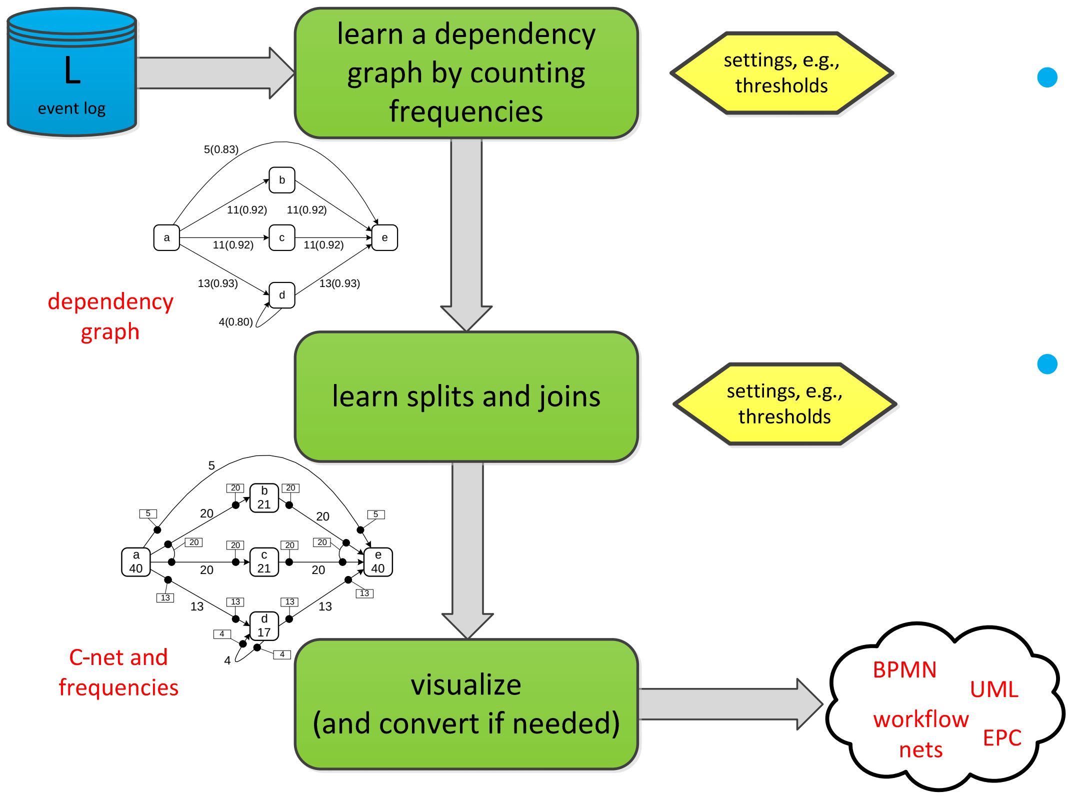
www.processmining.org



Technische Universiteit Eindhoven University of Technology

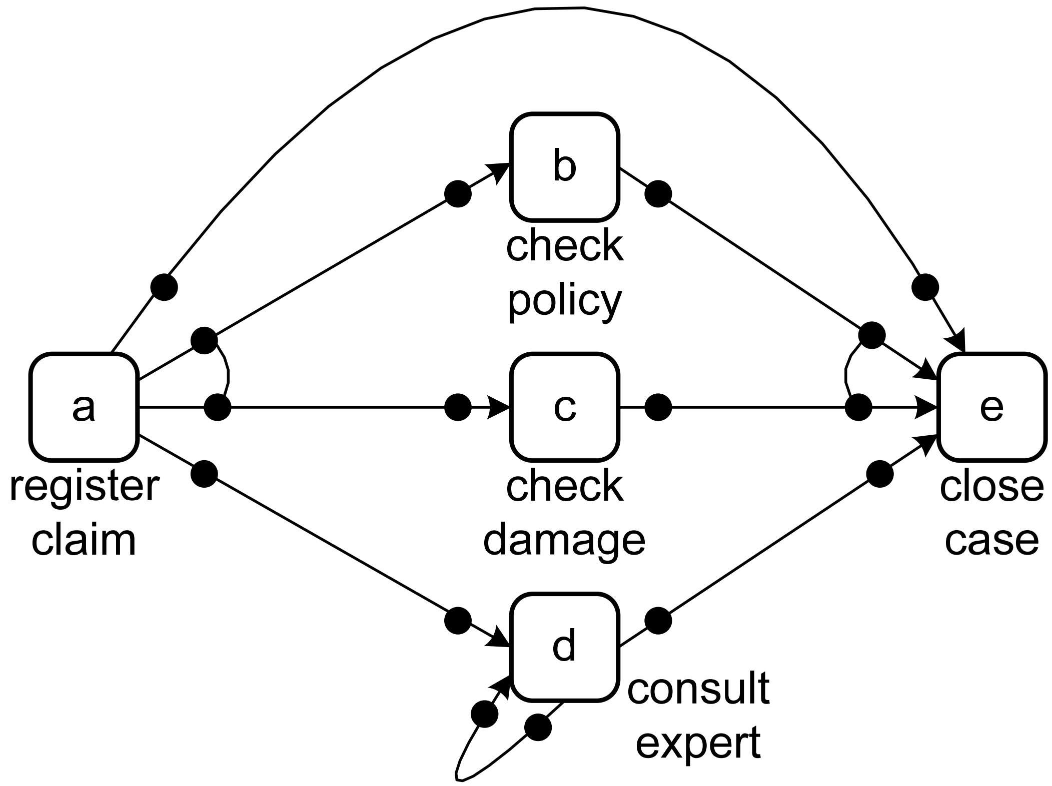
Where innovation starts

Heuristic mining: Two main phases

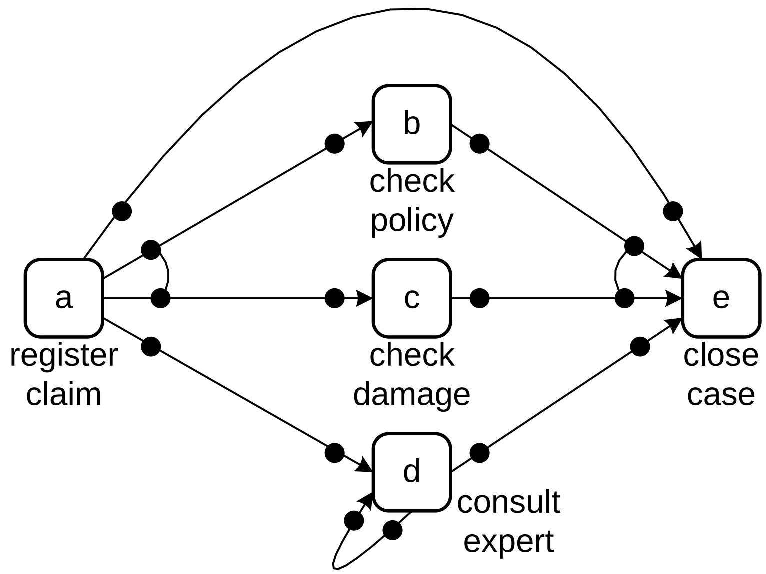


- Here we focus on learning **dependency graphs** (first phase).
- Inspired by heuristics miner (but many variations are possible).

Running example: C-net



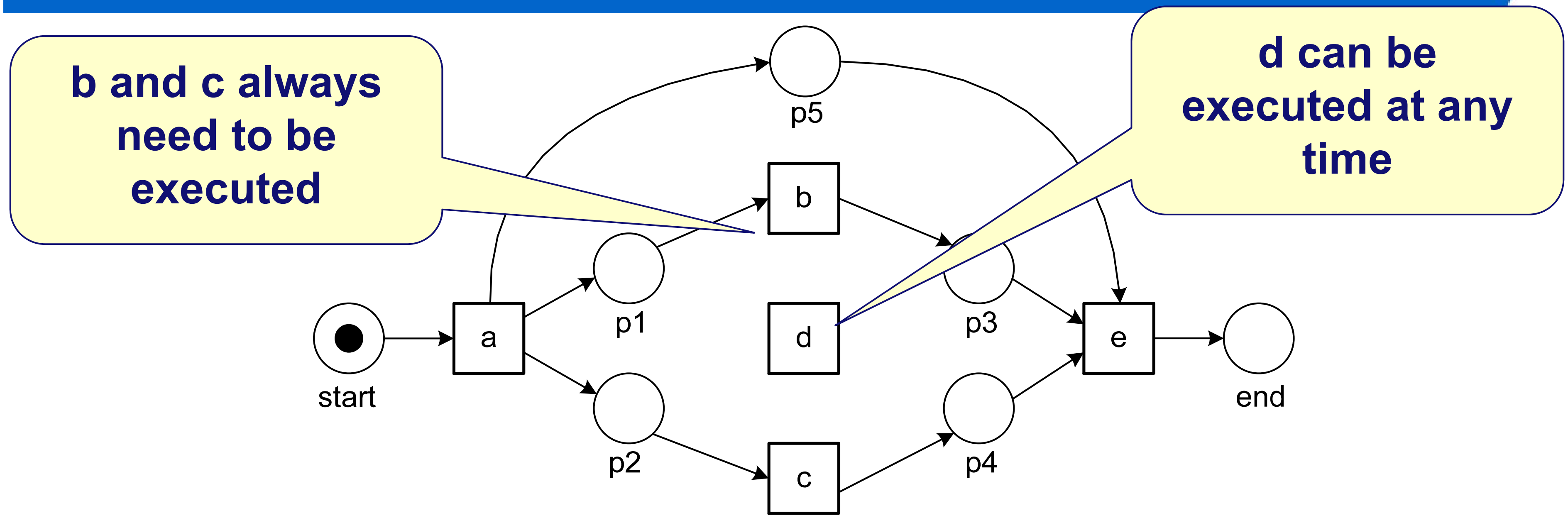
Example event log with some outliers



outliers (not in model)

$$L = [\langle a,e\rangle^{5}, \langle a,b,c,e\rangle^{10}, \langle a,c,b,e\rangle^{10} \boxed{, \langle a,b,e\rangle^{1}, \langle a,c,e\rangle^{1},} \\ \langle a,d,e\rangle^{10}, \langle a,d,d,e\rangle^{2}, \langle a,d,d,d,e\rangle^{1}]$$

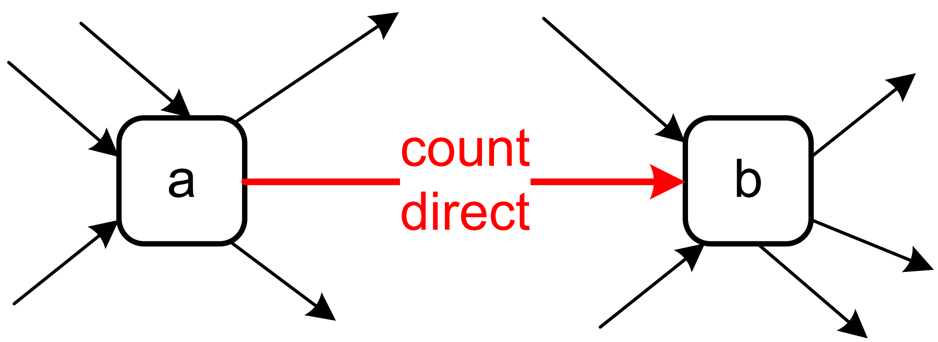
Problems of the Alpha algorithm with log



$$L = [\langle a,e\rangle^5, \langle a,b,c,e\rangle^{10}, \langle a,c,b,e\rangle^{10}, \langle a,b,e\rangle^1, \langle a,c,e\rangle^1,$$
$$\langle a,d,e\rangle^{10}, \langle a,d,d,e\rangle^2, \langle a,d,d,d,e\rangle^1]$$

Frequencies matter!

$$|a >_L b| = \sum_{\sigma \in L} L(\sigma) \times |\{1 \leq i < |\sigma| \mid \sigma(i) = a \wedge \sigma(i+1) = b\}|$$



Counting direct succession

information loss when frequencies are ignored

$$L = [\langle a,e\rangle^{5}, \langle a,b,c,e\rangle^{10}, \langle a,c,b,e\rangle^{10}, \langle a,b,e\rangle^{1}, \langle a, \ldots$$
$$\langle a,d,e\rangle^{10}, \langle a,d,d,e\rangle^{2}, \langle a,d,d,d,e\rangle^{1}]$$

$$|a >_L b| = \sum_{\sigma \in L} L(\sigma) \ldots \leq i < |\sigma| \mid \sigma(i) = a \wedge \sigma(i+1) = b\}|$$

$\lvert >_L \rvert$	a	b	c	d	e
a	0	11	11	true	5
b	0	0	10	false	11
c	0	10	0	false	11
d	0	0	0	true	13
e	0	0	0	false	0

Dependency measure
Taking into account concurrency

$$|a >_L b| = \sum_{\sigma \in L} L(\sigma) \times |\{1 \leq i < |\sigma| \mid \sigma(i) = a \wedge \sigma(i+1) = b\}|$$

direct succession

dependency measure

$|a \Rightarrow_L b|$ is the value of the dependency relation between a and b:

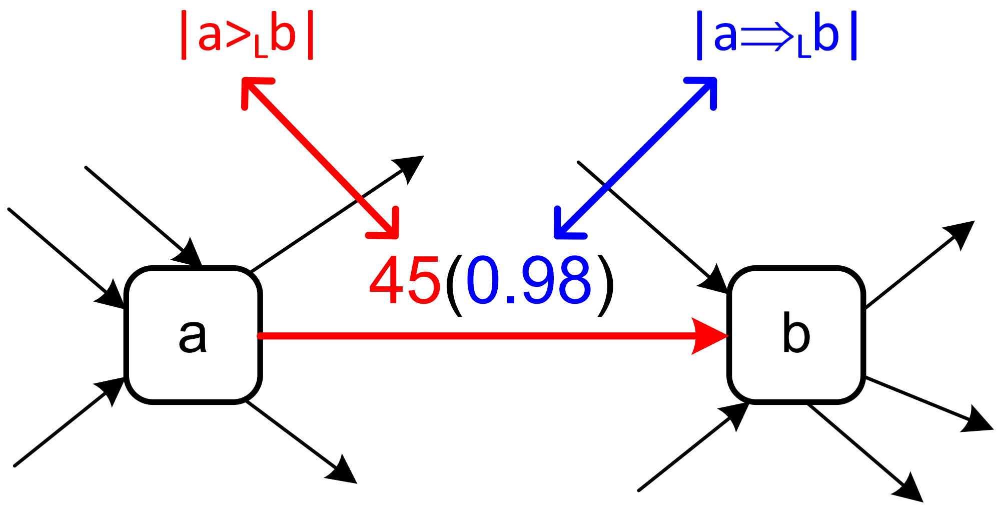
$$|a \Rightarrow_L b| = \begin{cases} \dfrac{|a >_L b| - |b >_L a|}{|a >_L b| + |b >_L a| + 1} & \text{if } a \neq b \\ \dfrac{|a >_L a|}{|a >_L a| + 1} & \text{if } a = b \end{cases}$$

Two values:
Direct succession and dependency

$$|a >_L b| = \sum_{\sigma \in L} L(\sigma) \times |\{1 \leq i < |\sigma| \mid \sigma(i) = a \wedge \sigma(i+1) = b\}|$$

$|a \Rightarrow_L b|$ is the value of the dependency relation between a and b:

$$|a \Rightarrow_L b| = \begin{cases} \dfrac{|a >_L b| - |b >_L a|}{|a >_L b| + |b >_L a| + 1} & \text{if } a \neq b \\ \dfrac{|a >_L a|}{|a >_L a| + 1} & \text{if } a = b \end{cases}$$



- **Both need to be above predefined thresholds!**
- **Otherwise, no causality!**

Sequence pattern

$|a \Rightarrow_L b|$ is the value of the dependency relation between a and b:

$$|a \Rightarrow_L b| = \begin{cases} \frac{|a >_L b| - |b >_L a|}{|a >_L b| + |b >_L a| + 1} & \text{if } a \neq b \\ \frac{|a >_L a|}{|a >_L a| + 1} & \text{if } a = b \end{cases}$$

both ($>_L$ and $\Rightarrow_L$) need to be above a threshold to be included

APPROVED

$|a >_L b| = 10$

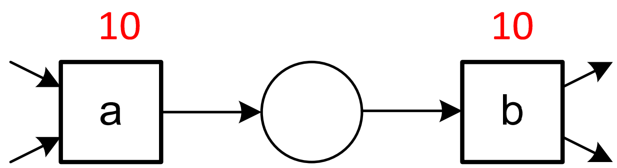
$|a \Rightarrow_L b| = 10/11$

$|b >_L a| = 0$

$|b \Rightarrow_L a| = -10/11$

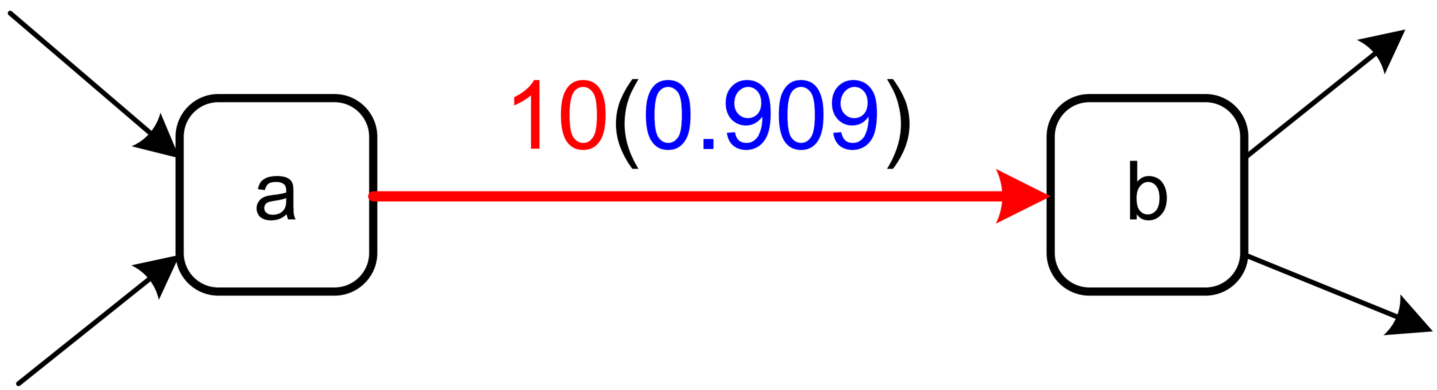
TU/e

Included arc (assuming thresholds ≥1 and ≥0.5)



$|a >_L b| = 10$ $\quad$ $|b >_L a| = 0$

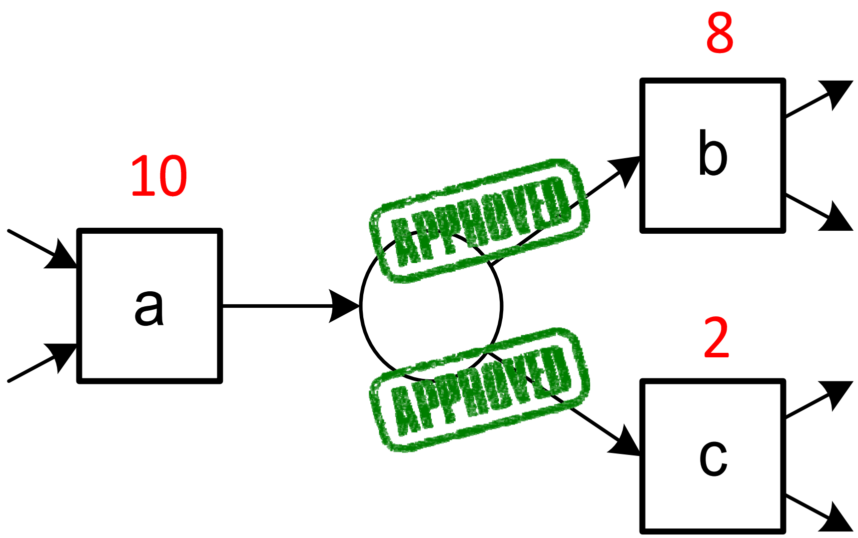
$|a \Rightarrow_L b| = 10/11$ $\quad$ $|b \Rightarrow_L a| = -10/11$



XOR-split pattern

$|a \Rightarrow_L b|$ is the value of the dependency relation between a and b:

$$|a \Rightarrow_L b| = \begin{cases} \frac{|a >_L b| - |b >_L a|}{|a >_L b| + |b >_L a| + 1} & \text{if } a \neq b \\ \frac{|a >_L a|}{|a >_L a| + 1} & \text{if } a = b \end{cases}$$



$|a >_L b| = 8$ $\quad$ $|b >_L a| = 0$

$|a \Rightarrow_L b| = 8/9$ $\quad$ $|b \Rightarrow_L a| = -8/9$

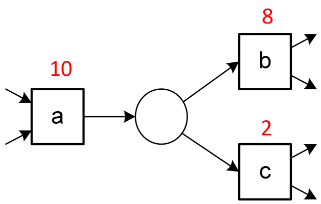
$|a >_L c| = 2$ $\quad$ $|b >_L c| = 0$

$|a \Rightarrow_L c| = 2/3$ $\quad$ $|b \Rightarrow_L c| = 0/1$

$|c >_L a| = 0$ $\quad$ $|c >_L b| = 0$

$|c \Rightarrow_L a| = -2/3$ $\quad$ $|c \Rightarrow_L b| = 0/1$

Included arcs (assuming thresholds ≥1 and ≥0.5)



$|a >_L b| = 8$ $|b >_L a| = 0$

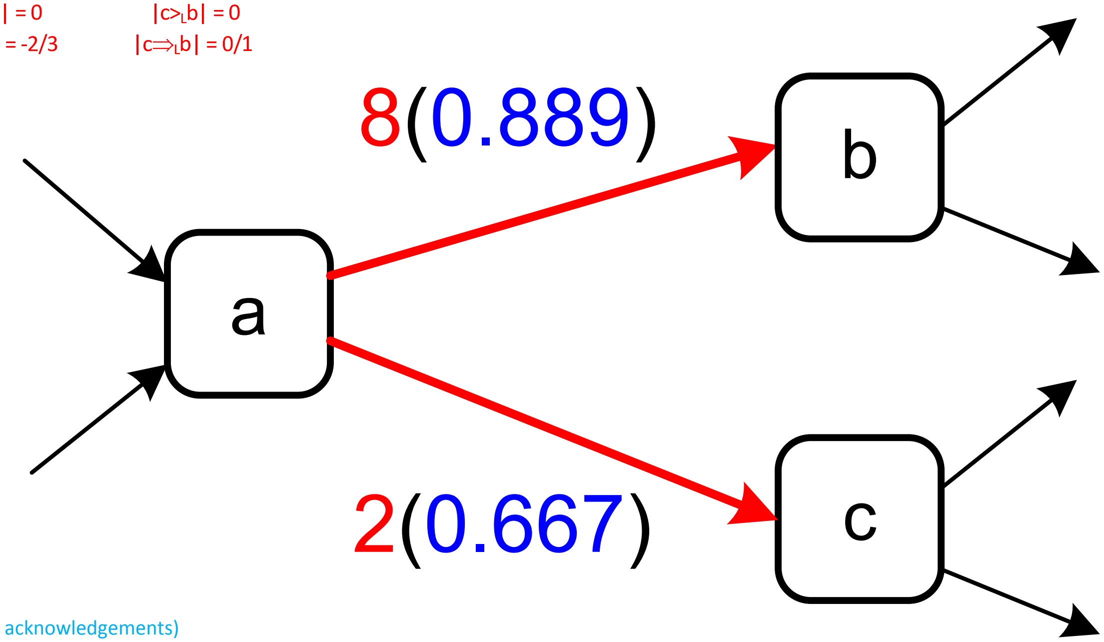
$|a \Rightarrow_L b| = 8/9$ $|b \Rightarrow_L a| = -8/9$

$|a >_L c| = 2$ $|b >_L c| = 0$

$|a \Rightarrow_L c| = 2/3$ $|b \Rightarrow_L c| = 0/1$

$|c >_L a| = 0$ $|c >_L b| = 0$

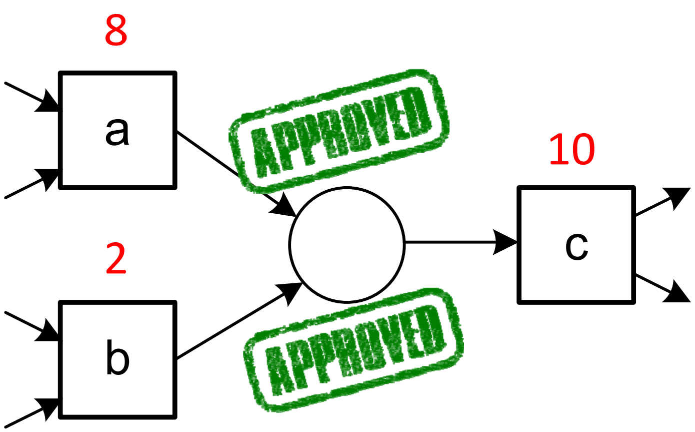
$|c \Rightarrow_L a| = -2/3$ $|c \Rightarrow_L b| = 0/1$



XOR-join pattern

$|a \Rightarrow_L b|$ is the value of the dependency relation between a and b:

$$|a \Rightarrow_L b| = \begin{cases} \dfrac{|a >_L b| - |b >_L a|}{|a >_L b| + |b >_L a| + 1} & \textit{if } a \neq b \\ \dfrac{|a >_L a|}{|a >_L a| + 1} & \textit{if } a = b \end{cases}$$



$|a >_L b| = 0$

$|a \Rightarrow_L b| = 0/1$

$|b >_L a| = 0$

$|b \Rightarrow_L a| = 0/1$

$|a >_L c| = 8$

$|a \Rightarrow_L c| = 8/9$

$|b >_L c| = 2$

$|b \Rightarrow_L c| = 2/3$

$|c >_L a| = 0$

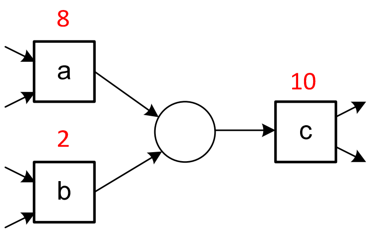
$|c \Rightarrow_L a| = -8/9$

$|c >_L b| = 0$

$|c \Rightarrow_L b| = -2/3$

TU/e

Included arcs (assuming thresholds ≥1 and ≥0.5)



$|a >_L b| = 0$
$|a \Rightarrow_L b| = 0/1$

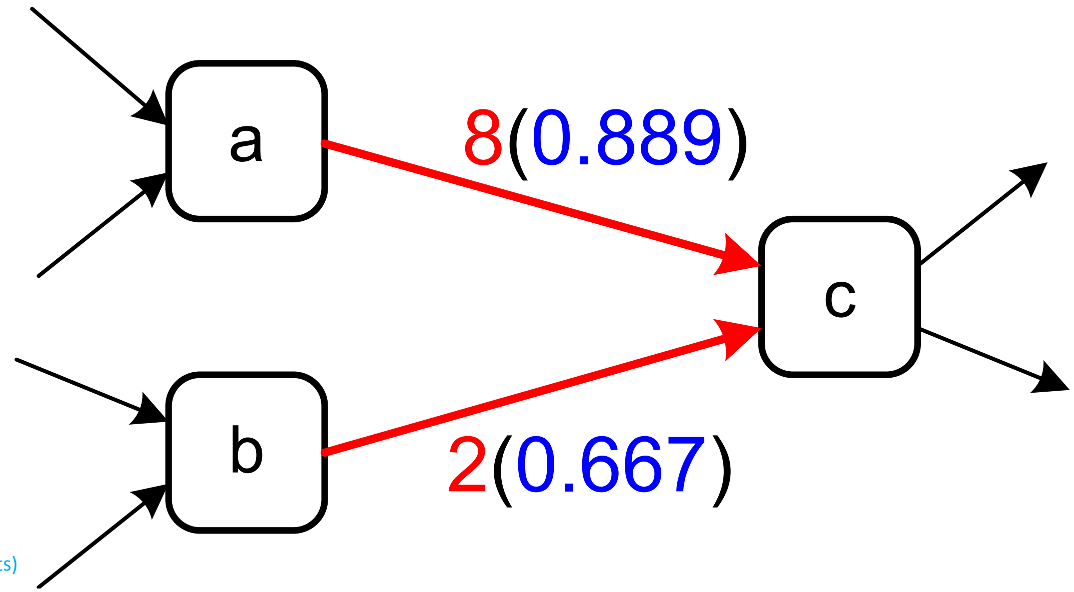
$|a >_L c| = 8$
$|a \Rightarrow_L c| = 8/9$

$|c >_L a| = 0$
$|c \Rightarrow_L a| = -8/9$

$|b >_L a| = 0$
$|b \Rightarrow_L a| = 0/1$

$|b >_L c| = 2$
$|b \Rightarrow_L c| = 2/3$

$|c >_L b| = 0$
$|c \Rightarrow_L b| = -2/3$

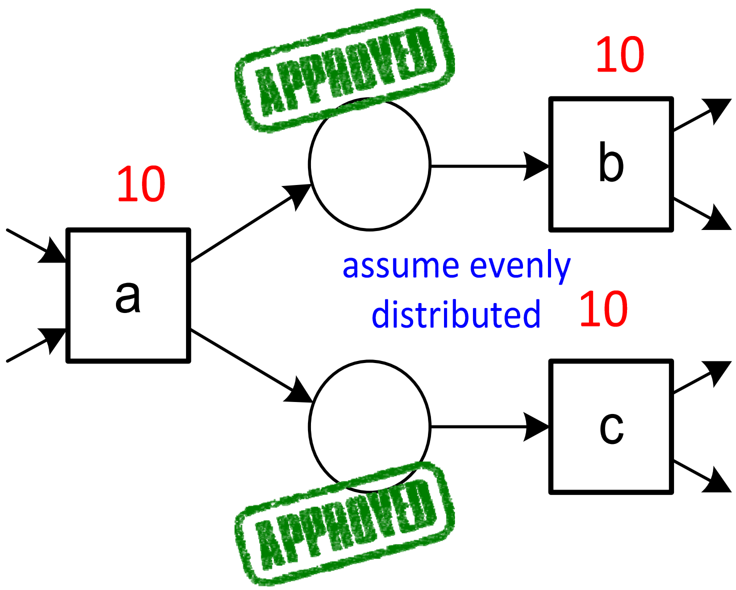


AND-split pattern

$|a \Rightarrow_L b|$ is the value of the dependency relation between a and b:

$$|a \Rightarrow_L b| = \begin{cases} \dfrac{|a >_L b| - |b >_L a|}{|a >_L b| + |b >_L a| + 1} & \text{if } a \neq b \\ \dfrac{|a >_L a|}{|a >_L a| + 1} & \text{if } a = b \end{cases}$$

illustrates why $>_L$ is not enough and $\Rightarrow_L$ is needed



APPROVED

$|a >_L b| = 5$

$|a \Rightarrow_L b| = 5/6$

$|b >_L a| = 0$

$|b \Rightarrow_L a| = -5/6$

APPROVED

$|a >_L c| = 5$

$|a \Rightarrow_L c| = 5/6$

$|b >_L c| = 5$!

$|b \Rightarrow_L c| = 0/11$

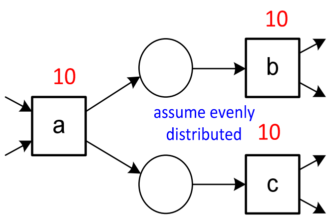
$|c >_L a| = 0$

$|c \Rightarrow_L a| = -5/6$

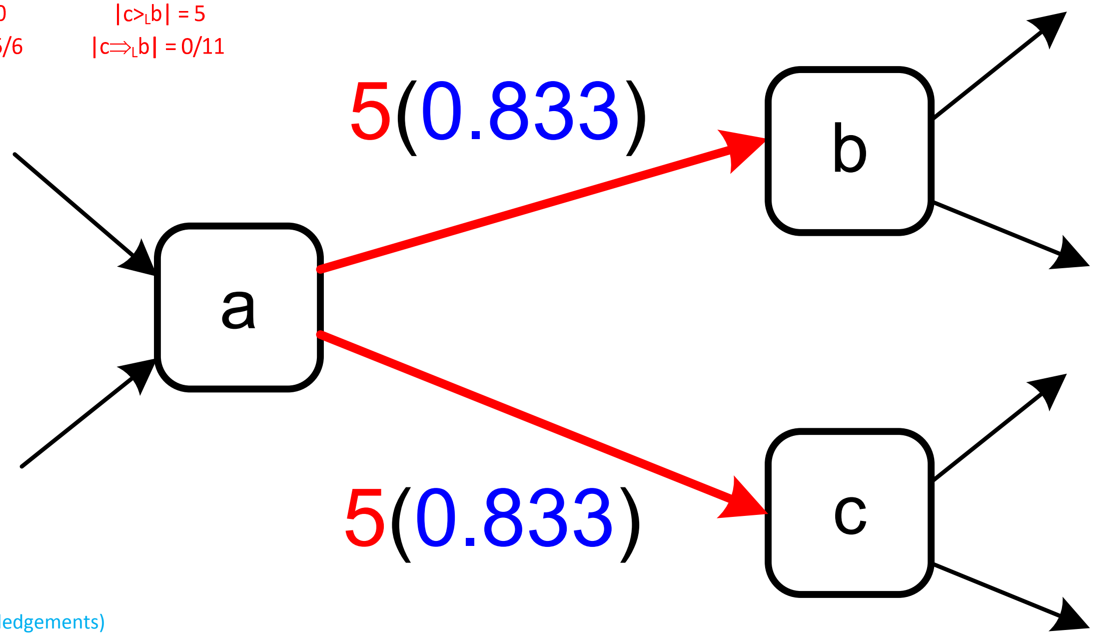
$|c >_L b| = 5$!

$|c \Rightarrow_L b| = 0/11$

Included arcs (assuming thresholds ≥1 and ≥0.5)



$\lvert a >_L b \rvert = 5$	$\lvert b >_L a \rvert = 0$
$\lvert a \Rightarrow_L b \rvert = 5/6$	$\lvert b \Rightarrow_L a \rvert = -5/6$
$\lvert a >_L c \rvert = 5$	$\lvert b >_L c \rvert = 5$
$\lvert a \Rightarrow_L c \rvert = 5/6$	$\lvert b \Rightarrow_L c \rvert = 0/11$
$\lvert c >_L a \rvert = 0$	$\lvert c >_L b \rvert = 5$
$\lvert c \Rightarrow_L a \rvert = -5/6$	$\lvert c \Rightarrow_L b \rvert = 0/11$

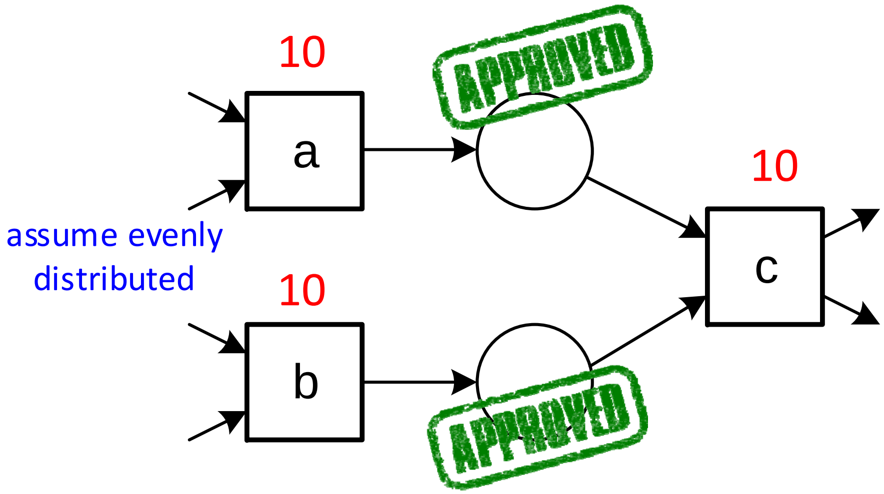


AND-join pattern

$|a \Rightarrow_L b|$ is the value of the dependency relation between a and b:

$$|a \Rightarrow_L b| = \begin{cases} \dfrac{|a >_L b| - |b >_L a|}{|a >_L b| + |b >_L a| + 1} & \textit{if } a \neq b \\ \dfrac{|a >_L a|}{|a >_L a| + 1} & \textit{if } a = b \end{cases}$$

illustrates why $>_L$ is not enough and $\Rightarrow_L$ is needed



$|a>_L b| = 5$!

$|a \Rightarrow_L b| = 0/11$ APPROVED

$|a>_L c| = 5$

$|a \Rightarrow_L c| = 5/6$

$|c>_L a| = 0$

$|c \Rightarrow_L a| = -5/6$

$|b>_L a| = 5$!

$|b \Rightarrow_L a| = 0/11$ APPROVED

$|b>_L c| = 5$

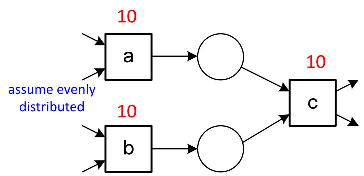
$|b \Rightarrow_L c| = 5/6$

$|c>_L b| = 0$

$|c \Rightarrow_L b| = -5/6$

TU/e

Included arcs (assuming thresholds ≥1 and ≥0.5)



$|a >_L b| = 5$

$|a \Rightarrow_L b| = 0/11$

$|b >_L a| = 5$

$|b \Rightarrow_L a| = 0/11$

$|a >_L c| = 5$

$|a \Rightarrow_L c| = 5/6$

$|b >_L c| = 5$

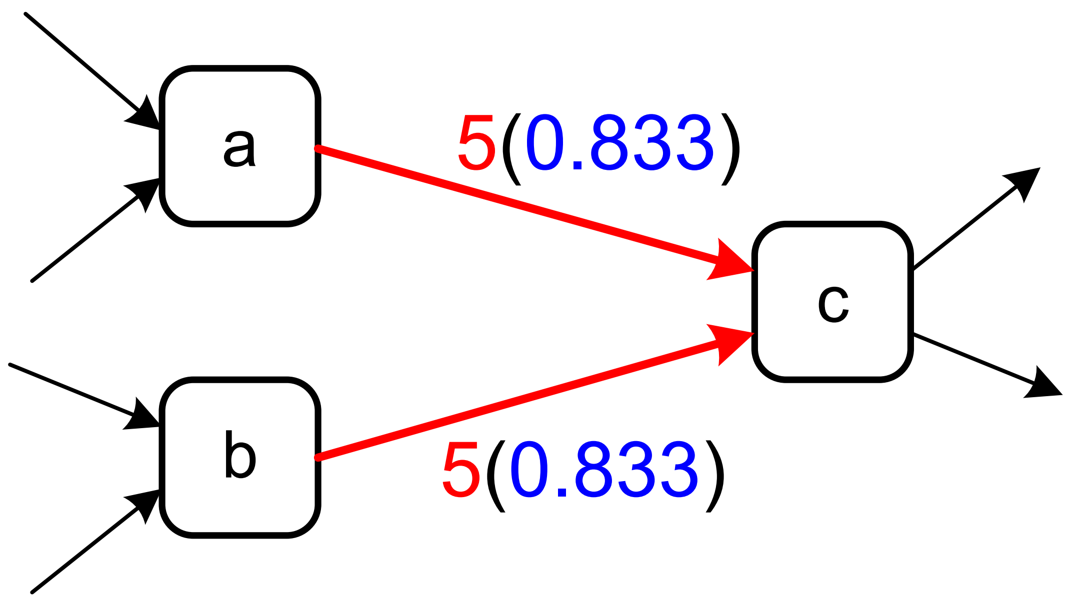
$|b \Rightarrow_L c| = 5/6$

$|c >_L a| = 0$

$|c \Rightarrow_L a| = -5/6$

$|c >_L b| = 0$

$|c \Rightarrow_L b| = 5/6$

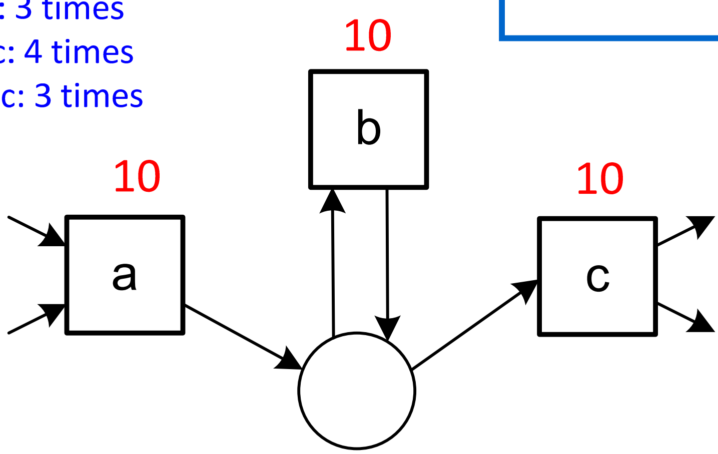


Loop pattern

$|a \Rightarrow_L b|$ is the value of the dependency relation between a and b:

$$|a \Rightarrow_L b| = \begin{cases} \dfrac{|a >_L b| - |b >_L a|}{|a >_L b| + |b >_L a| + 1} & \text{if } a \neq b \\ \dfrac{|a >_L a|}{|a >_L a| + 1} & \text{if } a = b \end{cases}$$

assume
ac: 3 times
abc: 4 times
abbc: 3 times



TU/e

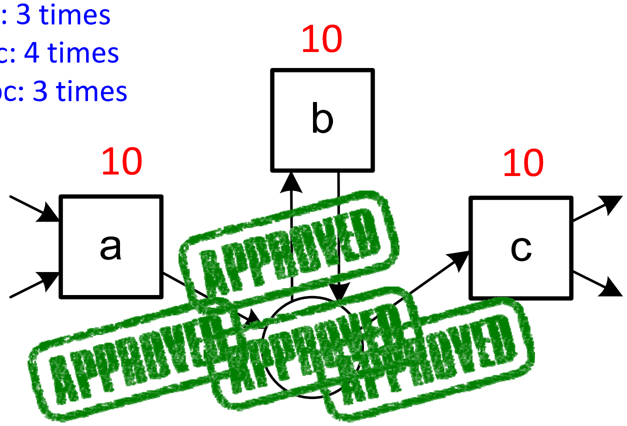
Loop pattern

$|a \Rightarrow_L b|$ is the value of the dependency relation between a and b:

$$|a \Rightarrow_L b| = \begin{cases} \dfrac{|a >_L b| - |b >_L a|}{|a >_L b| + |b >_L a| + 1} & \text{if } a \neq b \\ \dfrac{|a >_L a|}{|a >_L a| + 1} & \text{if } a = b \end{cases}$$

illustrates why self loops are handled differently (otherwise 0)

assume
ac: 3 times
abc: 4 times
abbc: 3 times



$|a>_L a| = 0$
$|a \Rightarrow_L a| = 0/1$

$|b>_L a| = 0$
$|b \Rightarrow_L a| = -7/8$

$|c>_L a| = 0$
$|c \Rightarrow_L a| = -3/4$

$|a>_L b| = 7$
$|a \Rightarrow_L b| = 7/8$

$|b>_L b| = 3$
$|b \Rightarrow_L b| = 3/4$

$|c>_L b| = 0$
$|c \Rightarrow_L b| = -7/8$

$|a>_L c| = 3$
$|a \Rightarrow_L c| = 3/4$

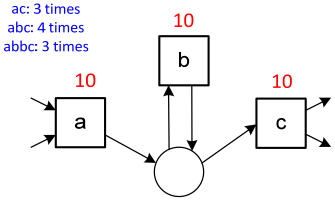
$|b>_L c| = 7$
$|b \Rightarrow_L c| = 7/8$

$|c>_L c| = 0$
$|c \Rightarrow_L c| = 0/1$

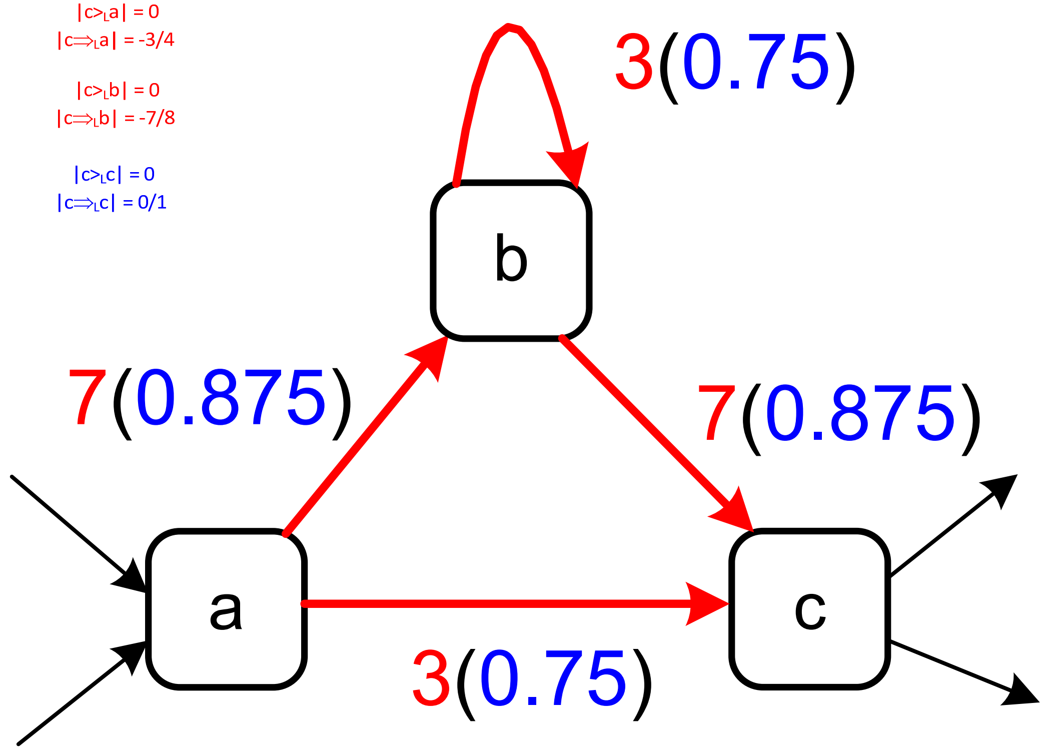
TU/e

Included arcs (assuming thresholds ≥1 and ≥0.5)

assume
ac: 3 times
abc: 4 times
abbc: 3 times



$\lvert a>_L a\rvert = 0$	$\lvert b>_L a\rvert = 0$	$\lvert c>_L a\rvert = 0$
$\lvert a\Rightarrow_L a\rvert = 0/1$	$\lvert b\Rightarrow_L a\rvert = -7/8$	$\lvert c\Rightarrow_L a\rvert = -3/4$
$\lvert a>_L b\rvert = 7$	$\lvert b>_L b\rvert = 3$	$\lvert c>_L b\rvert = 0$
$\lvert a\Rightarrow_L b\rvert = 7/8$	$\lvert b\Rightarrow_L b\rvert = 3/4$	$\lvert c\Rightarrow_L b\rvert = -7/8$
$\lvert a>_L c\rvert = 3$	$\lvert b>_L c\rvert = 7$	$\lvert c>_L c\rvert = 0$
$\lvert a\Rightarrow_L c\rvert = 3/4$	$\lvert b\Rightarrow_L c\rvert = 7/8$	$\lvert c\Rightarrow_L c\rvert = 0/1$



Example revisited

$$L = [\langle a,e\rangle^5, \langle a,b,c,e\rangle^{10}, \langle a,c,b,e\rangle^{10}, \langle a,b,e\rangle^1, \langle a,c,e\rangle^1,$$
$$\langle a,d,e\rangle^{10}, \langle a,d,d,e\rangle^2, \langle a,d,d,d,e\rangle^1]$$

$\vert >_L \vert$	a	b	c	d	e
a	0	11	11	13	5
b	0	0	10	0	11
c	0	10	0	0	11
d	0	0	0	4	13
e	0	0	0	0	0

Question: What are $|a \Rightarrow_L b|$ and $|d \Rightarrow_L d|$?

$$L = [\langle a,e\rangle^5, \langle a,b,c,e\rangle^{10}, \langle a,c,b,e\rangle^{10}, \langle a,b,e\rangle^1, \langle a,c,e\rangle^1,$$
$$\langle a,d,e\rangle^{10}, \langle a,d,d,e\rangle^2, \langle a,d,d,d,e\rangle^1]$$

$\lvert >_L \rvert$	a	b	c	d	e
a	0	11	11	13	5
b	0	0	10	0	11
c	0	10	0	0	11
d	0	0	0	4	13
e	0	0	0	0	0

Compute the dependency measures $|a \Rightarrow_L b|$ **and** $|d \Rightarrow_L d|$

Dependency measures computed for all activity pairs

$\lvert \Rightarrow_L \rvert$	a	b	c	d	e
a	$\frac{0}{0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$	$\frac{11-0}{11+0+1} = 0.92$	$\frac{13-0}{13+0+1} = 0.93$	$\frac{5-0}{5+0+1} = 0.83$
b	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0}{0+1} = 0$	$\frac{10-10}{10+10+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$
c	$\frac{0-11}{0+11+1} = -0.92$	$\frac{10-10}{10+10+1} = 0$	$\frac{0}{0+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$
d	$\frac{0-13}{0+13+1} = -0.93$	$\frac{0-0}{0+0+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{4}{4+1} = 0.80$	$\frac{13-0}{13+0+1} = 0.93$
e	$\frac{0-5}{0+5+1} = -0.83$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0-13}{0+13+1} = -0.93$	$\frac{0}{0+1} = 0$

Two example values: $|a \Rightarrow_L b|$ and $|d \Rightarrow_L d|$

$\|\Rightarrow_L\|$	a	b	c	d	e
a	$\frac{0}{0+1}=0$	$\boxed{\frac{11-0}{11+0+1}=0.92}$	$\frac{11-0}{11+0+1}=0.92$	$\frac{13-0}{13+0+1}=0.93$	$\frac{5-0}{5+0+1}=0.83$
b	$\frac{0-11}{0+11+1}=-0.92$	$\frac{0}{0+1}=0$	$\frac{10-10}{10+10+1}=0$	$\frac{0-0}{0+0+1}=0$	$\frac{11-0}{11+0+1}=0.92$
c	$\frac{0-11}{0+11+1}=-0.92$	$\frac{10-10}{10+10+1}=0$	$\frac{0}{0+1}=0$	$\frac{0-0}{0+0+1}=0$	$\frac{11-0}{11+0+1}=0.92$
				$\boxed{\frac{4}{4+1}=0.80}$	$\frac{13-0}{13+0+1}=0.93$
				$\frac{0-13}{0+13+1}=-0.93$	$\frac{0}{0+1}=0$

$\|>_L\|$	a	b	c	d	e
a	0	$\boxed{11}$	11	13	5
b	$\boxed{0}$	0	10	0	11
c	0	10	0	0	11
d	0	0	0	$\boxed{4}$	13
e	0	0	0	0	0

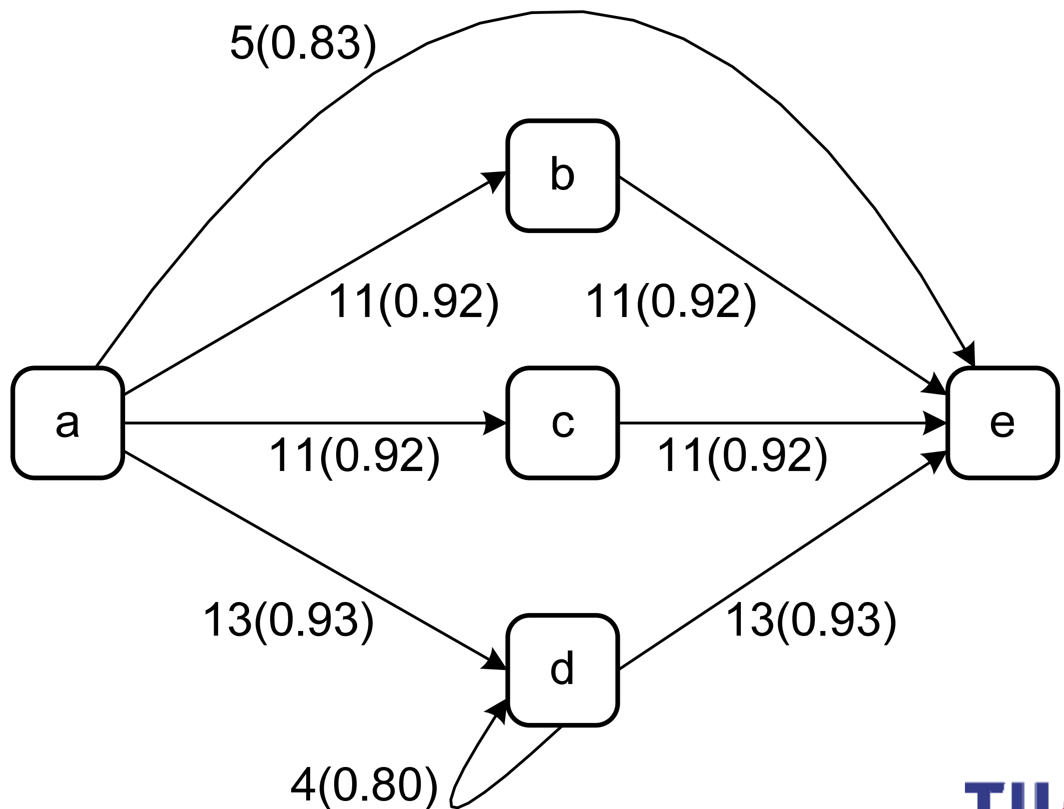
$|a \Rightarrow_L b|$ is the value of the dependency relation between a and b:

$$|a \Rightarrow_L b| = \begin{cases} \dfrac{|a >_L b| - |b >_L a|}{|a >_L b| + |b >_L a| + 1} & \text{if } a \neq b \\ \dfrac{|a >_L a|}{|a >_L a| + 1} & \text{if } a = b \end{cases}$$

Dependency graph using a lower threshold
(at least 2 direct successions and a dependency of at least 0.7)

$\vert \Rightarrow_L \vert$	a	b	c	d	e
a	$\frac{0}{0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$	$\frac{11-0}{11+0+1} = 0.92$	$\frac{13-0}{13+0+1} = 0.93$	$\frac{5-0}{5+0+1} = 0.83$
b	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0}{0+1} = 0$	$\frac{10-10}{10+10+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$
c	$\frac{0-11}{0+11+1} = -0.92$	$\frac{10-10}{10+10+1} = 0$	$\frac{0}{0+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$
d	$\frac{0-13}{0+13+1} = -0.93$	$\frac{0-0}{0+0+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{4}{4+1} = 0.80$	$\frac{13-0}{13+0+1} = 0.93$
e	$\frac{0-5}{0+5+1} = -0.83$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0-13}{0+13+1} = -0.93$	$\frac{0}{0+1} = 0$

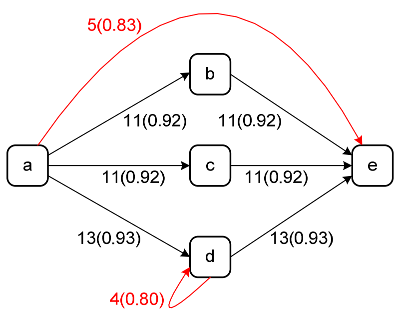
$\vert >_L \vert$	a	b	c	d	e
a	0	11	11	13	5
b	0	0	10	0	11
c	0	10	0	0	11
d	0	0	0	4	13
e	0	0	0	0	0

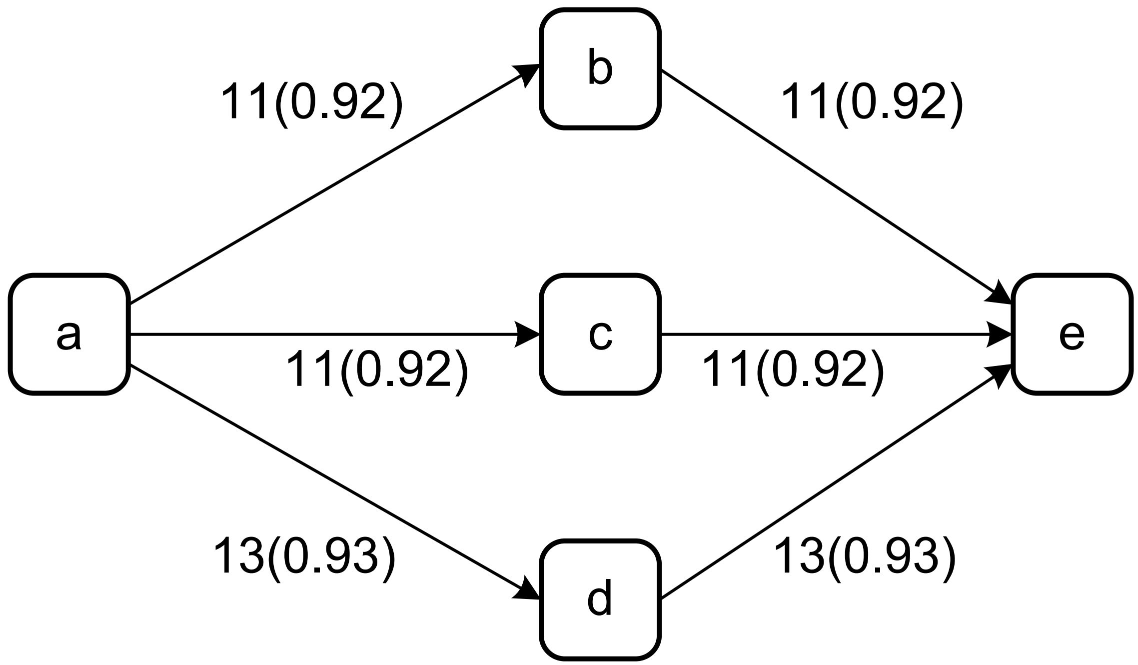


Dependency graph using a higher threshold
(at least 5 direct successions and a dependency of at least 0.9)

$\vert\Rightarrow_L\vert$	a	b	c	d	e
a	$\frac{0}{0+1}=0$	$\frac{11-0}{11+0+1}=0.92$	$\frac{11-0}{11+0+1}=0.92$	$\frac{13-0}{13+0+1}=0.93$	$\frac{5-0}{5+0+1}=0.83$
b	$\frac{0-11}{0+11+1}=-0.92$	$\frac{0}{0+1}=0$	$\frac{10-10}{10+10+1}=0$	$\frac{0-0}{0+0+1}=0$	$\frac{11-0}{11+0+1}=0.92$
c	$\frac{0-11}{0+11+1}=-0.92$	$\frac{10-10}{10+10+1}=0$	$\frac{0}{0+1}=0$	$\frac{0-0}{0+0+1}=0$	$\frac{11-0}{11+0+1}=0.92$
d	$\frac{0-13}{0+13+1}=-0.93$	$\frac{0-0}{0+0+1}=0$	$\frac{0-0}{0+0+1}=0$	$\frac{4}{4+1}=0.80$	$\frac{13-0}{13+0+1}=0.93$
e	$\frac{0-5}{0+5+1}=-0.83$	$\frac{0-11}{0+11+1}=-0.92$	$\frac{0-11}{0+11+1}=-0.92$	$\frac{0-13}{0+13+1}=-0.93$	$\frac{0}{0+1}=0$

$\vert>_L\vert$	a	b	c	d	e
a	0	11	11	13	5
b	0	0	10	0	11
c	0	10	0	0	11
d	0	0	0	4	13
e	0	0	0	0	0





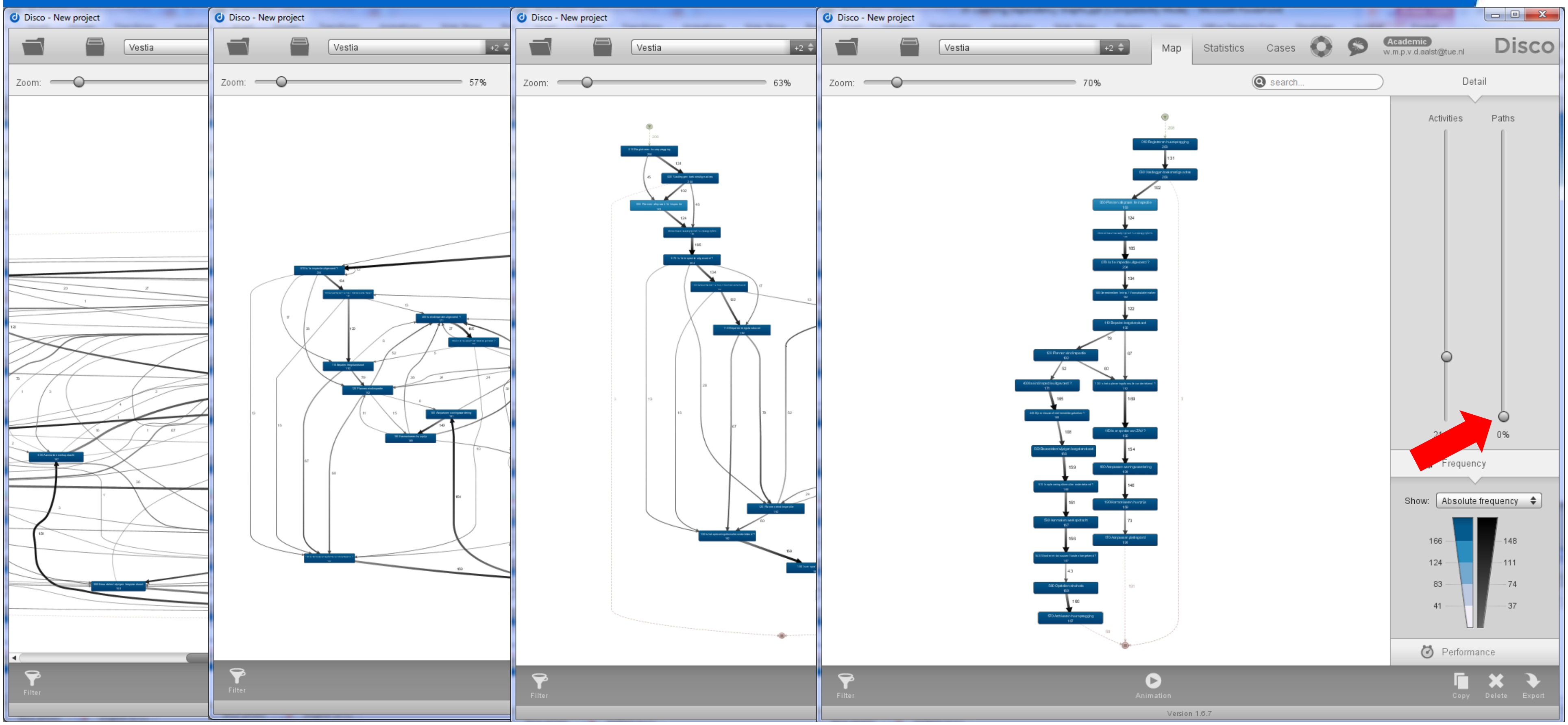
Computing the dependency graph

1. Set thresholds for the minimal number of direct successions and the dependency measure.
2. Count direct successions.
3. Compute dependency measures.
4. Draw dependency graph while including only arcs that meet both thresholds.

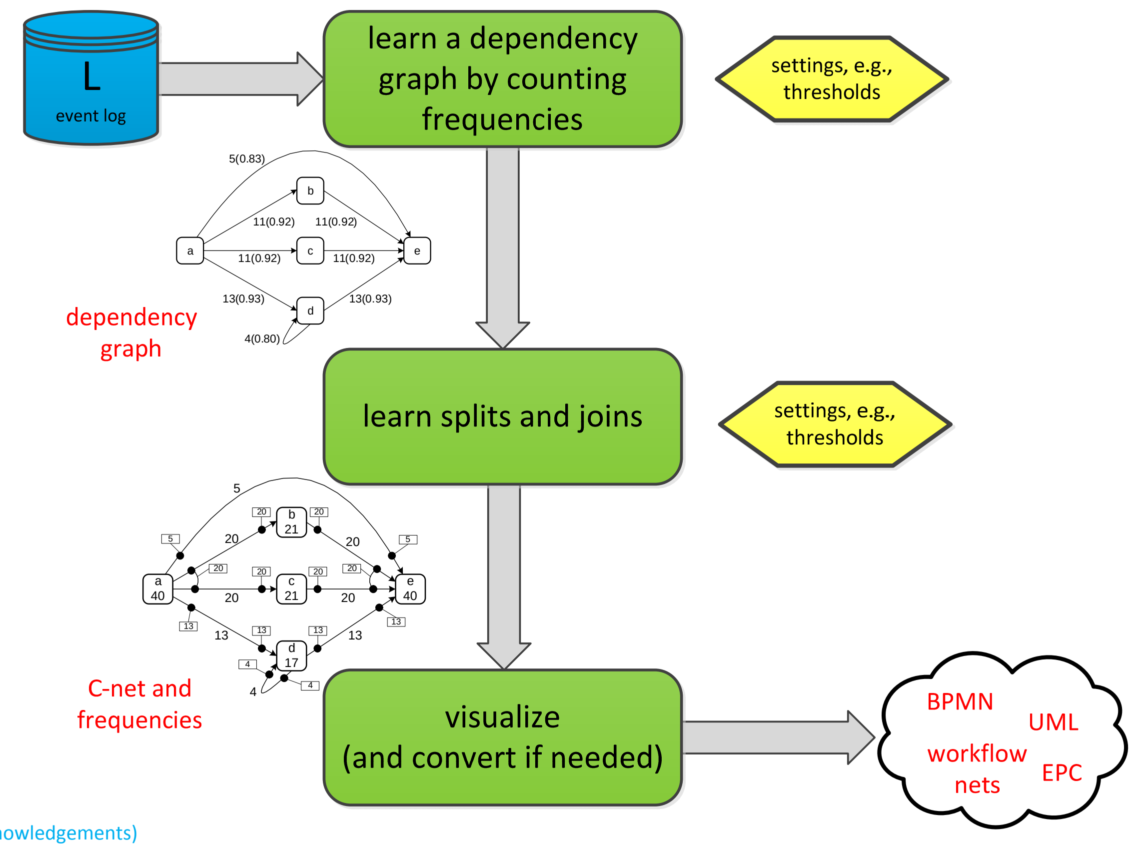
Practice doing this yourself on small example logs!

Disco

(different kind of dependency graphs, but idea of thresholds is similar)



Next step: Learn splits and joins



Part I: Preliminaries

Chapter 1
Introduction

Chapter 2
Process Modeling and Analysis

Chapter 3
Data Mining

Part II: From Event Logs to Process Models

Chapter 4
Getting the Data

Chapter 5
Process Discovery: An Introduction

Chapter 6
Advanced Process Discover[illegible]niques

Part III: Beyond Process Discovery

Chapter 7
Conformance Checking

Chapter 8
Mining Additional Perspectives

Chapter 9
Operational Support

Part IV: Putting Process Mining to Work

Chapter 10
Tool Support

Chapter 11
Analyzing “Lasagna Processes”

Chapter 12
Analyzing “Spaghetti Processes”

Part V: Reflection

Chapter 13
Cartography and Navigation

Chapter 14
Epilogue