How Similar Are Two Hospitals:
Using Probability And Geometry For Comparison

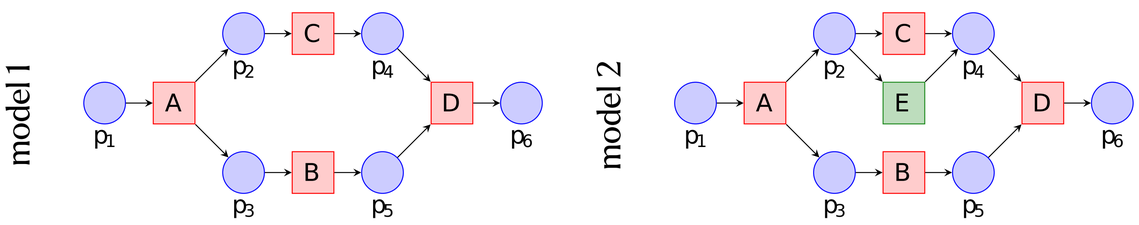
Antony R. Lee, P. Tiňo (University of Birmingham)
I. B. Styles (Queens University Belfast)

our aim

To propose a rigorous way to compare processes, using the structure of their models and the frequencies of observed traces

to begin, why?

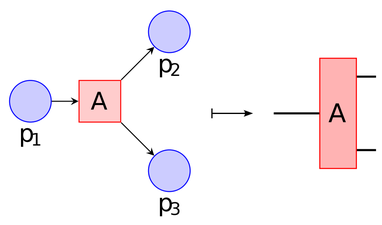
Comparing hospitals (or any healthcare process) is a central part of many national strategies for improving healthcare



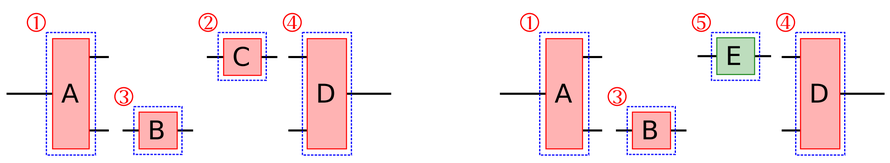
But what properties of a model should we compare?

break things up, for probabilities

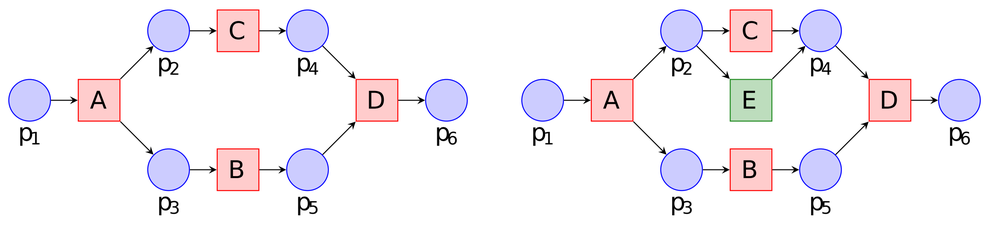
Property 1 (Structure) decompose process models into compositional elements to identify their common structure [1]



Across all models, pick a consistent sequence of element compositions



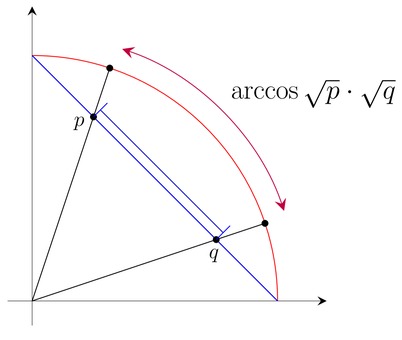
Each model is represented by a set of distinct sequences



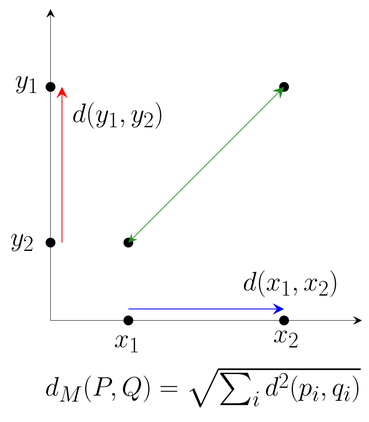
Key idea: the sequences can be used to develop probabilistic processes [2]

principled comparisons, using geometry

Probabilities are constrained vectors that impose a specific geometry (sum to 1), for which there's a corresponding arc-length based distance [3]



$$d(p,q) = \arccos \sqrt{p} \cdot \sqrt{q}$$



$$d_M(P,Q) = \sqrt{\textstyle\sum_i} d^2(p_i, q_i)$$

Probabilis processes can be represented by matrices, which are interpreted as collections of probabilities. Each arc-length distance of a pair of probabilities can be combined to give a matrix distance

Key idea: use the geometry of probabilities to compare models

putting it together, a distance

For two processes represented by stochastic matrices $P = (p_1 \dots, p_n)$ $Q = (q_1 \dots, q_n)$ the matrix distance between them is [4]

$$d_M(P,Q) = \sqrt{\textstyle\sum_i} \arccos^2 \sqrt{p_i} \cdot \sqrt{q_i}$$

an example

Property 2 (Frequencies) stochastic processes can be modelled with Probability Prefix Automata

model 1

$$L_1 = [\langle ABCD\rangle^{10}, \langle ACBD\rangle^{12}] \to [(1,3,2,4)^{22}]$$

model 2

$$L_2 = [\langle ABCD\rangle^{10}, \langle ACBD\rangle^{12}, \langle ABED\rangle^{100}, \langle AEBD\rangle^{110}] \to [(1,3,2,4)^{22}, (1,3,5,4)^{210}]$$

Read sequences one symbol at a time. For each position, look at all possible prefixes, in other words:

$$(1,3,2,4) \to \{1, 13, 132, 1324\} \qquad (1,3,5,4) \to \{1, 13, 135, 1354\}$$

Count how often each symbol follows every prefix. Turn these counts into probabilities for each prefix–symbol pair

$13 \to 132 : \frac{22}{232} \approx 0.1$
$13 \to 135 : \frac{210}{232} \approx 0.9$

For instance (model 2): if 13 has been observed, either 2 or 5 will be next with probability 0.1 or 0.9 respectively

Build automaton states for each prefix and add transitions with their probabilities, and convert to their matrix form for comparison

model 1

ε: [1: 1.0]
|- 1: [3: 1.0]
| |- 13: [2: 1.0]
| | |- 132: [4: 1.0]
| | | |- 1324: [ε: 1.0]
| | |- 135: [ε: 1.0]
| | | |- 1354: [ε: 1.0]

	ϵ	1	13	132	1324	135	1354
ϵ		1.0					
1			1.0				
13				1.0			
132					1.0		
1324	1.0						
135	1.0						
354	1.0						

model 2

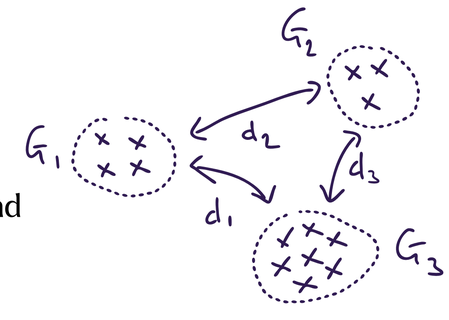
ε: [1: 1.0]
|- 1: [3: 1.0]
| |- 13: [2: 0.1, 5: 0.9]
| | |- 132: [4: 1.0]
| | | |- 1324: [ε: 1.0]
| | |- 135: [4: 1.0]
| | | |- 1354: [ε: 1.0]

	ϵ	1	13	132	1324	135	1354
ϵ		1.0					
1			1.0				
13				0.9		0.1	
132					1.0		
1324	1.0						
135							1.0
1354	1.0						

Key idea: distance is how far apart two things are, or how similar they are

lastly, and what?

Similarity, defined as smaller distances between processes, can be used to group hospitals/ departments/ teams etc



hospitals between G1 and G2 are different, why?

hospitals in G3 are similar, is this useful?

Most(?) process mining notations can be rewritten using the compositional approach (string diagrams)

Probability-geometric links provide a robust distance metric for comparing how close processes are

Decomposing models reveals their shared structure and enables probability assignment to observed traces

The distance function is efficient, and clustering and other grouping techniques can leverage this distance for model analysis

refs and deets