Privacy-Preserving Process Mining:

Towards the new European General Data Protection Regulation

Edgar Batista[1] and Agusti Solanas[2]

[1] SIMPPLE, S.L.
C. Joan Maragall 1A
43003 Tarragona, Catalonia, Spain
edgar.batista@simpple.com

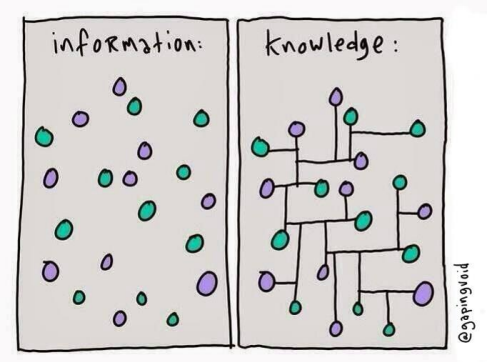
[2] Smart Health Research Group
Universitat Rovira i Virgili
Av. Paisos Catalans, 26
43007 Tarragona, Catalonia, Spain
agusti.solanas@urv.cat

Outline

1. Process Mining
2. Process Mining in Healthcare: Challenges & Opportunities
3. Privacy-Preserving Process Mining
4. Case study
5. Conclusions

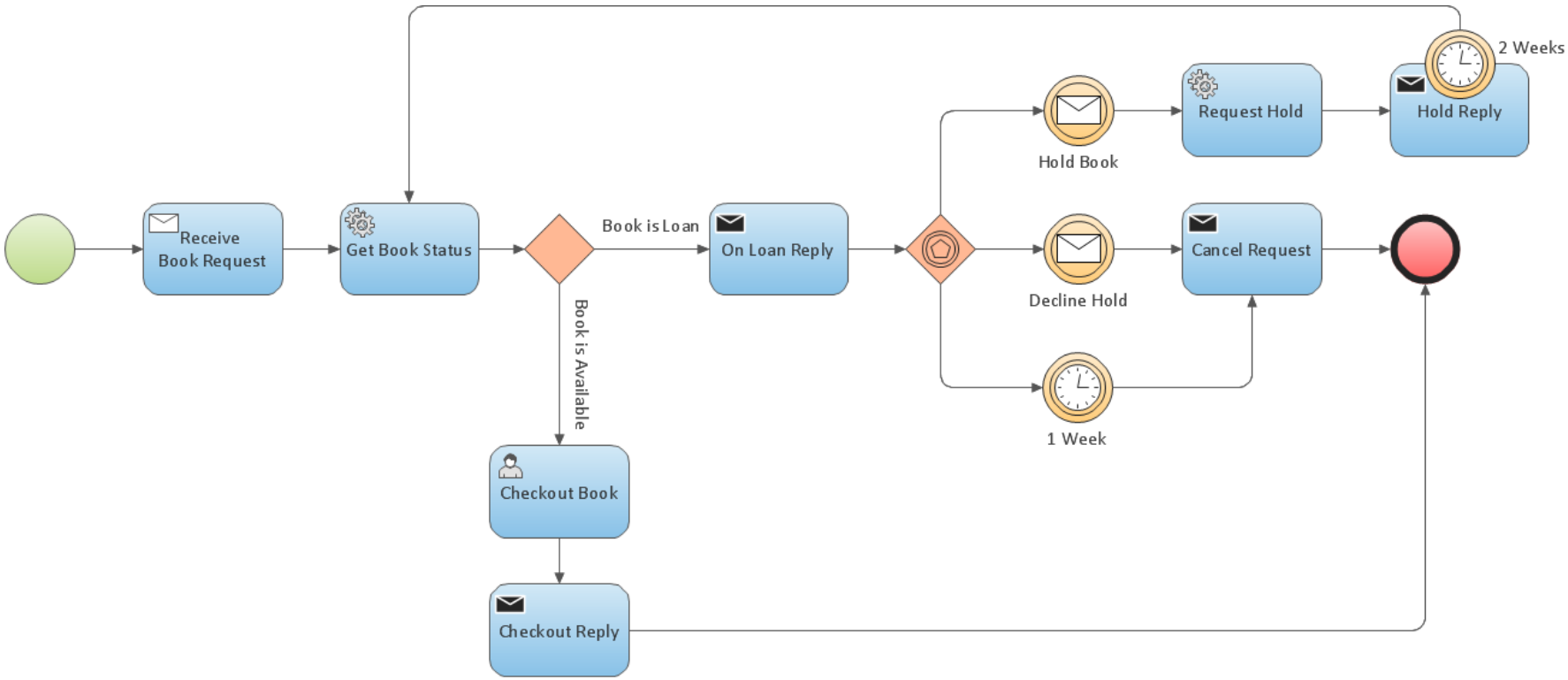
1. Process Mining

- Organizations deal with multiple **information systems**
  - MIS, DSS, Data Warehouses, ERP, GIS…
- Store as much **information** as possible to extract added-value **knowledge** and make better **decisions**
  - *"Knowledge is power"*
- **Business processes** play an important role in today's information systems
  - *"From data-aware to process-aware"*



1. Process Mining

- A business process is a **set of activities** aiming at accomplishing a certain **organizational goal**



1. Process Mining

- Existing techniques to monitor the execution of processes are based on **events**
  - Each event contains the activity executed, timestamp, different identifiers (*e.g.* resources),…
  - Events are stored in **event log files**

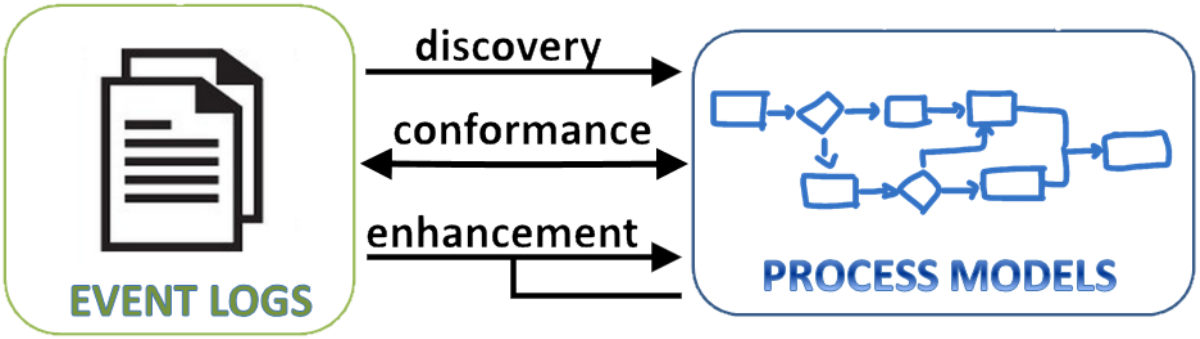
case id	event id	properties				
		timestamp	activity	resource	cost	…
1	35654423	30-12-2010:11.02	register request	Pete	50	…
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	…
	35654425	05-01-2011:15.12	check ticket	Mike	100	…
	35654426	06-01-2011:11.18	decide	Sara	200	…
	35654427	07-01-2011:14.24	reject request	Pete	200	…
2	35654483	30-12-2010:11.32	register request	Mike	50	…
	35654485	30-12-2010:12.12	check ticket	Mike	100	…
	35654487	30-12-2010:14.16	examine casually	Pete	400	…
	35654488	05-01-2011:11.22	decide	Sara	200	…
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	…

1. Process Mining

- **Process mining** is a research field aiming at discovering, monitoring and improving real business processes by extracting knowledge from the event logs available in organizational information systems

- Advantages:
  - Optimization of resources
  - Identification of bottlenecks
  - Detection of hidden dependencies
  - Better decision-makings for future improvements

1. Process Mining



a) **Discovery**: Produce process models from event logs without any a-priori information

b) **Conformance**: Verify the alignment between an existing process model with an event log of the same process

c) **Enhancement**: Extend or improve existing process models using the information stored in event log files from that processes

Outline

1. Process Mining
2. Process Mining in Healthcare: Challenges & Opportunities
3. Privacy-Preserving Process Mining
4. Case study
5. Conclusions

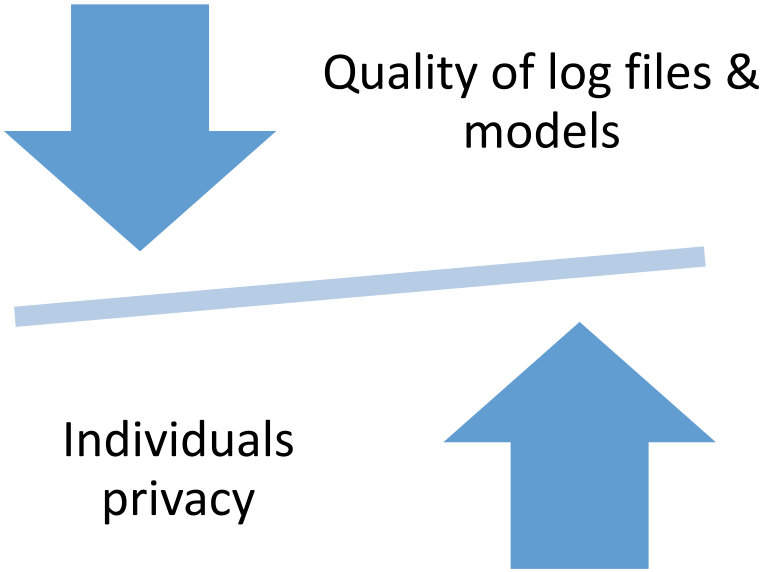
2. Process Mining in Healthcare: Challenges & Opportunities

- Healthcare event log files may contain **personal data**
  - Especially, **sensitive data** (patients/doctors identifiers, health conditions, treatments, diseases…)
- Careful management to **guarantee individuals privacy**
- **Distorting data** (*e.g.* micro-aggregation) to prevent the disclosure of sensitive data and avoid re-identification
  - Especially when subcontracting process mining services to **3rd parties**!

2. Process Mining in Healthcare: Challenges & Opportunities

- Process mining relies on the **quality of log files**
  - Trade-off: improve privacy but not reducing quality

- Research on process mining in healthcare uses raw data to obtain accurate and realistic views of healthcare processes, **BUT**…



2. Process Mining in Healthcare: Challenges & Opportunities

EU GDPR **strengthens** the protection of personal data, specially those referring to sensitive data, for all individuals within the EU

2. Process Mining in Healthcare: Challenges & Opportunities

- How GDPR affects to process mining analyses
  - **Art. 2**: *"This Regulation applies to the processing of personal data wholly or partly by automated means [...]"*
  - **Art. 4**: "*Personal data means any information relating to an identifiable natural person, [...] who can be identified, directly or indirectly, by reference to an identifier such as a name, an identification number, location data, an online identifier [...]*"
  - **Art. 5**: "*Personal data shall be: (a) processed lawfully, fairly and in a transparent manner [...], (b) collected for explicit and legitimate purposes [...], (c) adequate, relevant and limited to what is necessary [...], (f) processed in a manner that ensures appropriate security of the personal data [...]*"

2. Process Mining in Healthcare: Challenges & Opportunities

- How GDPR affects to process mining analyses
  - **Art. 6**: "*Processing shall be lawful [...]: (a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes*"
  - **Art. 7**: "*[...] the request for consent shall be presented in a manner which is clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language*"
  - **Art. 9**: "*Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data, data concerning health, sex life or sexual orientation shall be prohibited, [... unless] has given explicit consent to the processing of those personal data for one or more specified purposes*"

2. Process Mining in Healthcare: Challenges & Opportunities

- How GDPR affects to process mining analyses
  - **Art. 25**: *"implement appropriate technical and organisational measures, such as pseudonymisation, […] and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects"*
  - **Art. 32**: *"[…] implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk […]: (a) the pseudonymisation and encryption of personal data; (b) the ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and services […]"*

https://www.i-scoop.eu/gdpr/gdpr-personal-data-identifiers-pseudonymous-information/
https://iapp.org/news/a/looking-to-comply-with-gdpr-heres-a-primer-on-anonymization-and-pseudonymization
https://iapp.org/news/a/top-10-operational-impacts-of-the-gdpr-part-8-pseudonymization/

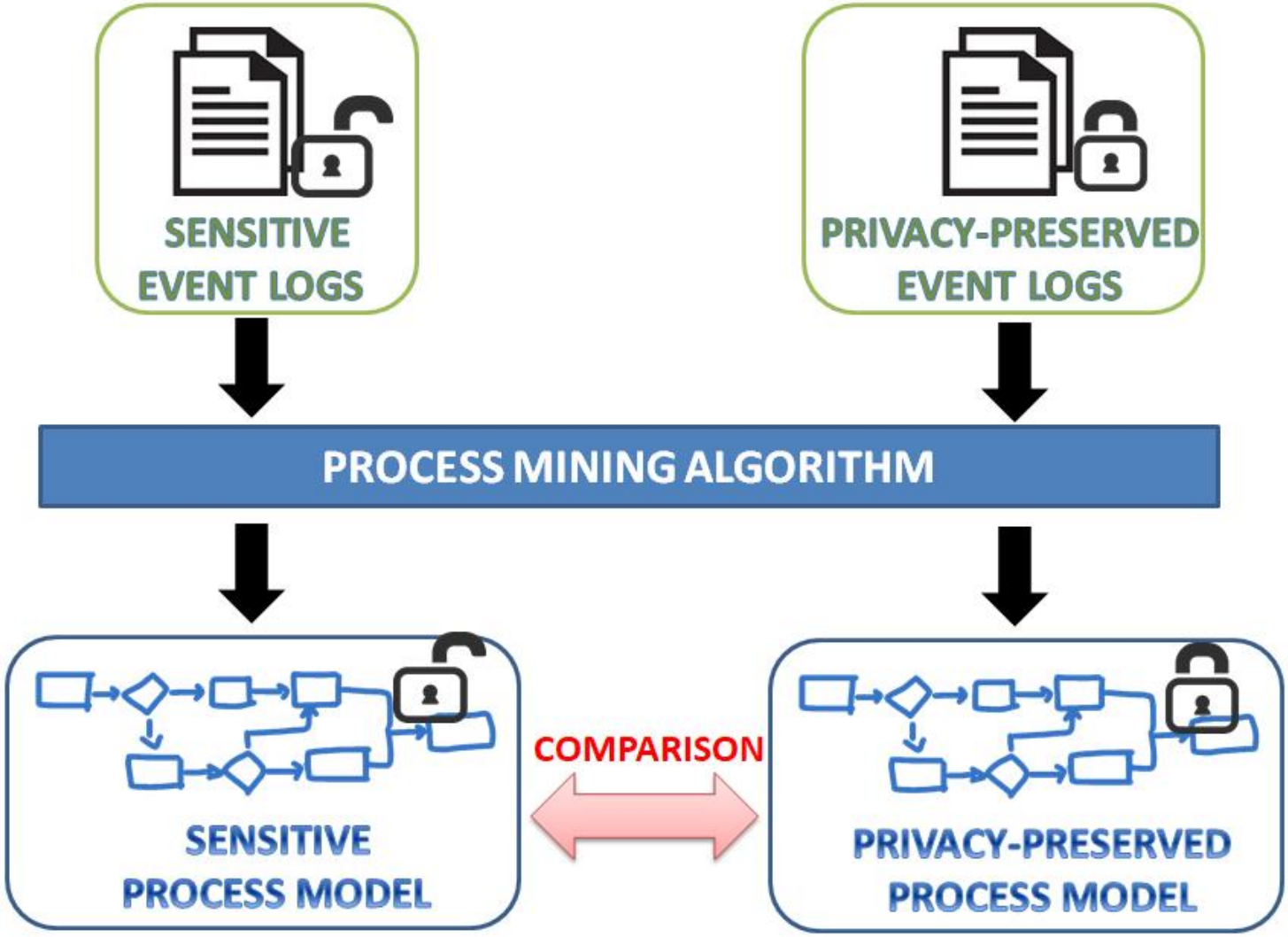
Outline

1. Process Mining
2. Process Mining in Healthcare: Challenges & Opportunities
3. Privacy-Preserving Process Mining
4. Case study
5. Conclusions

3. Privacy-Preserving Process Mining

- Existing works do not use any privacy-preserving technique (realistic views)
- No research on assessing the effectiveness of current process mining methods with proper privacy-preserved datasets
- **GOAL**: Studying how process models differ when they are generated from raw events or privacy-preserved events

3. Privacy-Preserving Process Mining



Outline

1. Process Mining
2. Process Mining in Healthcare: Challenges & Opportunities
3. Privacy-Preserving Process Mining
4. Case study
5. Conclusions

4. Case Study

- Event log file (≈ 130k events) of requests/petitions of doctors in a hospital
- **What we focus on?**: How doctors behave?
  - (Transitions between) actions that doctors perform during patient treatments
  - 1 doctor : 1 process behaviour (graph)
- Not limited to this question

4. Case Study

	id integer	episode text	doctor text	patient text	e_action text	e_time timestamp without time zone
1	1	1400012|T	AGUCULREY	300181	4:petMa	2015-04-09 10:04:52
2	2	1400012|T	AGUCULREY	300181	3:petMa	2015-04-09 10:05:05
3	719	1528749|T	LOPSANSAN	293744	4:petRx	2015-06-25 19:59:10
4	720	1528749|T	LOPSANSAN	293744	3:petRx	2015-06-25 19:59:14
5	721	1528749|T	LOPSANSAN	293744	5:petRx	2015-06-25 19:59:20
6	722	1528749|T	LOPSANSAN	293744	4:petRx	2015-06-25 20:08:33
7	723	1528749|T	LOPSANSAN	293744	3:petRx	2015-06-25 20:08:40
8	724	1528749|T	LOPSANSAN	293744	5:petRx	2015-06-25 20:08:42
9	762	1528798|T	HERTOBSEB	1604304	4:petInter	2015-06-26 08:13:21
10	1038	1529717|T	LOPSANSAN	1856922	4:petRx	2015-07-01 11:25:04
11	1039	1529717|T	LOPSANSAN	1856922	3:petRx	2015-07-01 11:25:12
12	1040	1529717|T	LOPSANSAN	1856922	5:petRx	2015-07-01 11:25:15
13	1041	1529717|T	LOPSANSAN	1856922	3:petRx	2015-07-01 11:28:59
14	1042	1529717|T	LOPSANSAN	1856922	5:petRx	2015-07-01 11:29:01

4. Case Study

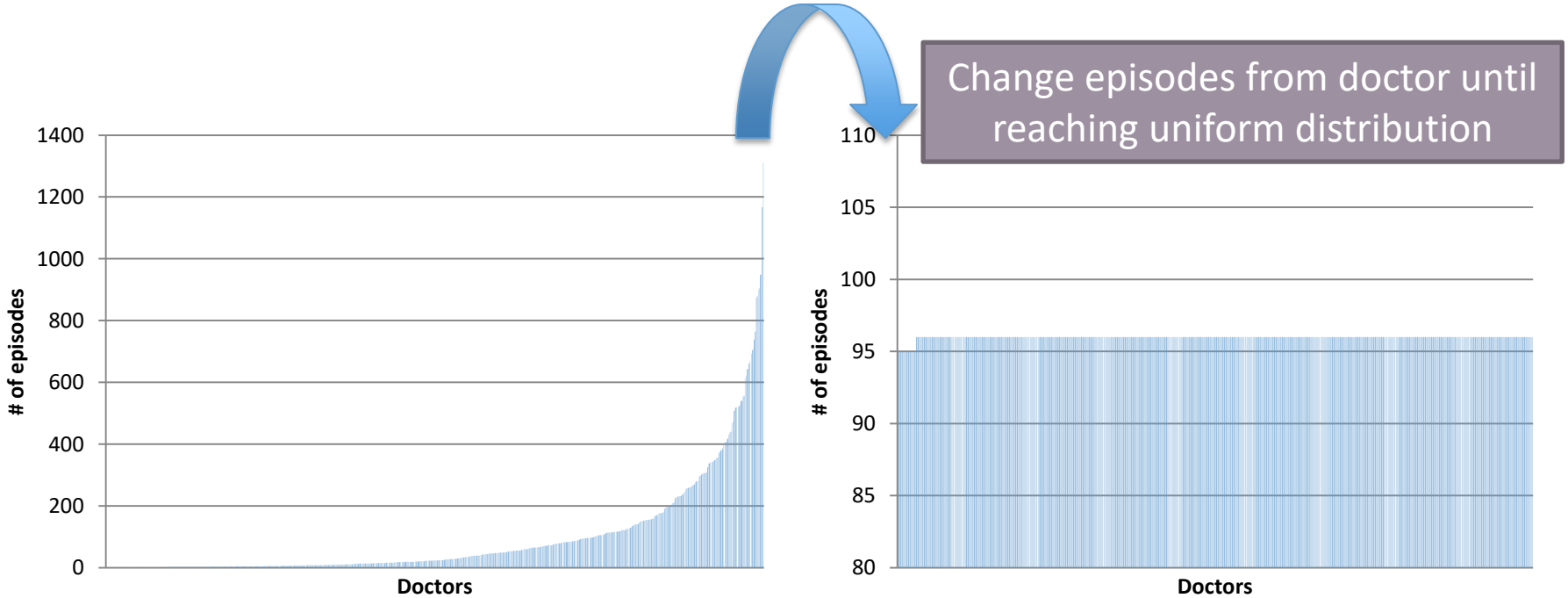
- **Problem**: Some doctors have much more episodes than others (different distribution)

- **Adversary model**:
  - Assumed scenario: Distribution of doctors can be known (# of consultation hours…)
  - Profiling of doctors: Individualization
  - Goal: Identify the behavior of each doctor

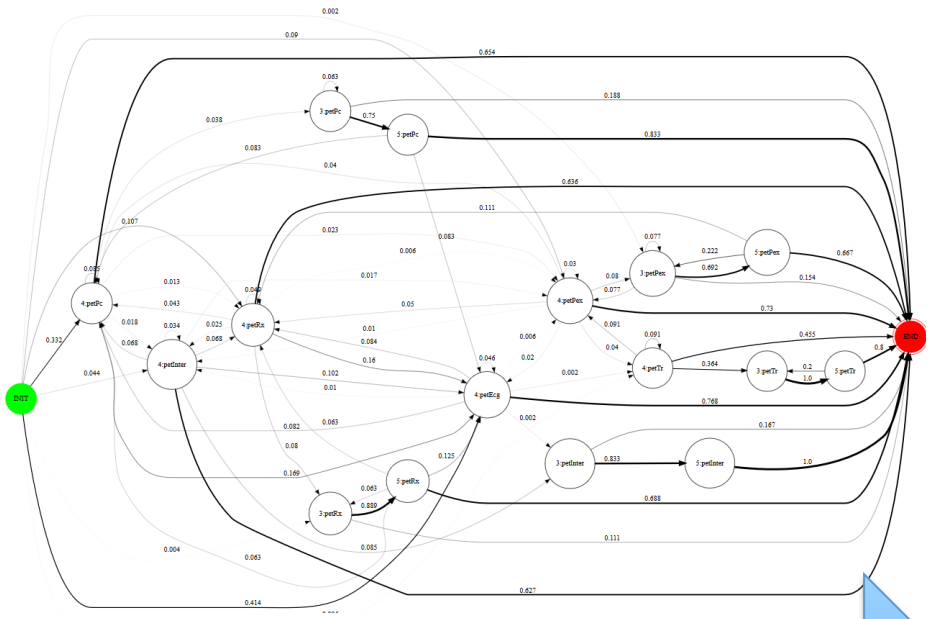
- **Privacy-preserving solution**:
  - Pseudonymization is not enough (no affects distribution)
  - Uniform the distribution of episodes per doctor as much as possible (hide the doctor's episodes) → Change doctor ID
  - Prevent re-identification of doctors

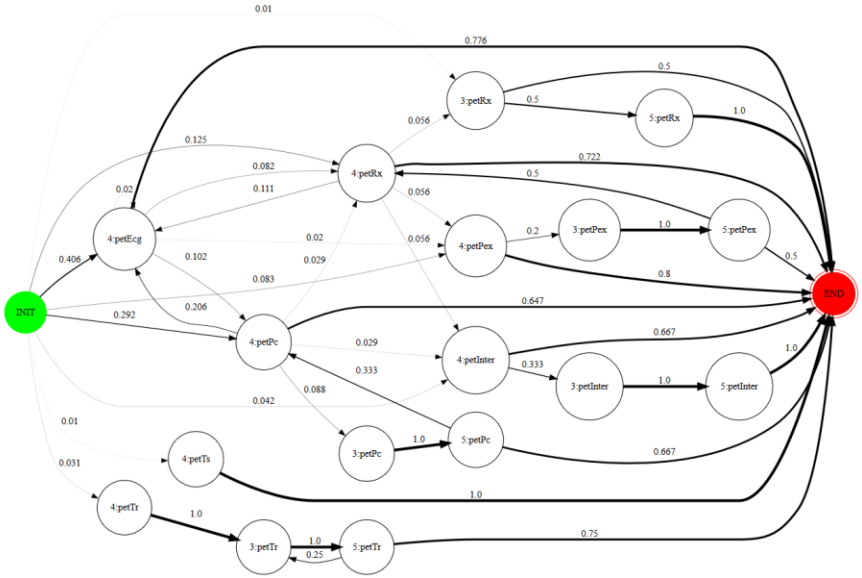
4. Case Study

- Distribution of episodes per doctor
  - 648 doctors
  - ≈ 62k episodes
  - #episodes per doctor: [1, 1312] => [95, 96]



4. Case Study



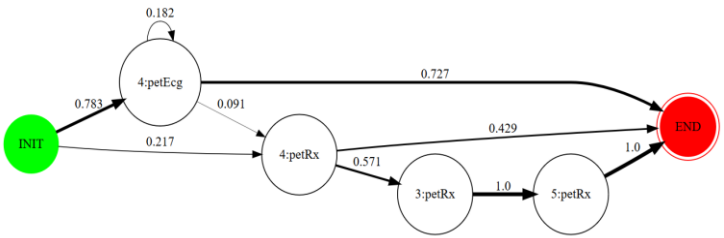


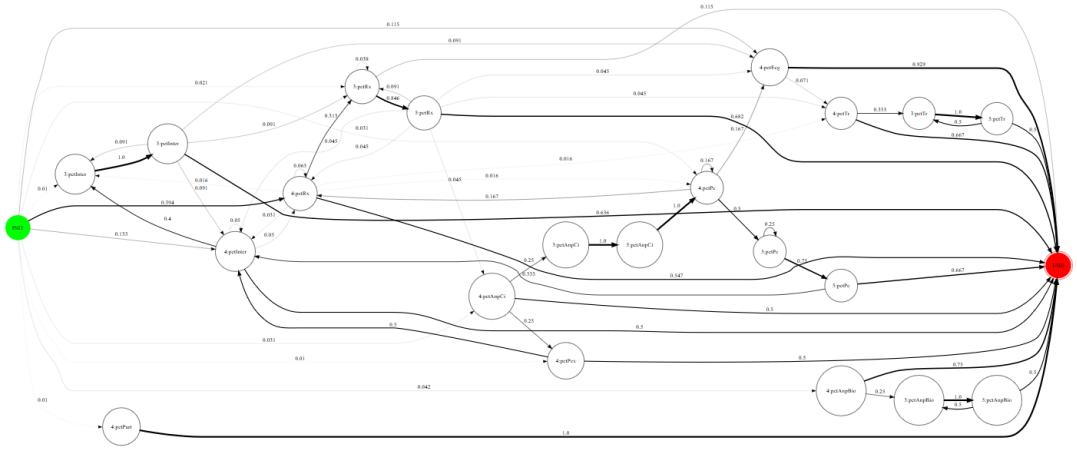
Simplification
Little inf. loss

- 900 episodes
- 18 nodes
- 75 edges
- Most frequent paths: <INIT, 4:petPC, END>, <INIT, 4:petEcg, END>, <INIT, 4:petRx, END>
- Node degree: 8,3 ± 4,7

- 96 episodes
- 19 nodes
- 43 edges: 41 (same) + 2 (new)
- Most frequent paths: <INIT, 4:petEcg, END> <INIT, 4:petPC, END> <INIT, 4:petRx, END>
- Node degree: 4,5 ± 2,9

4. Case Study



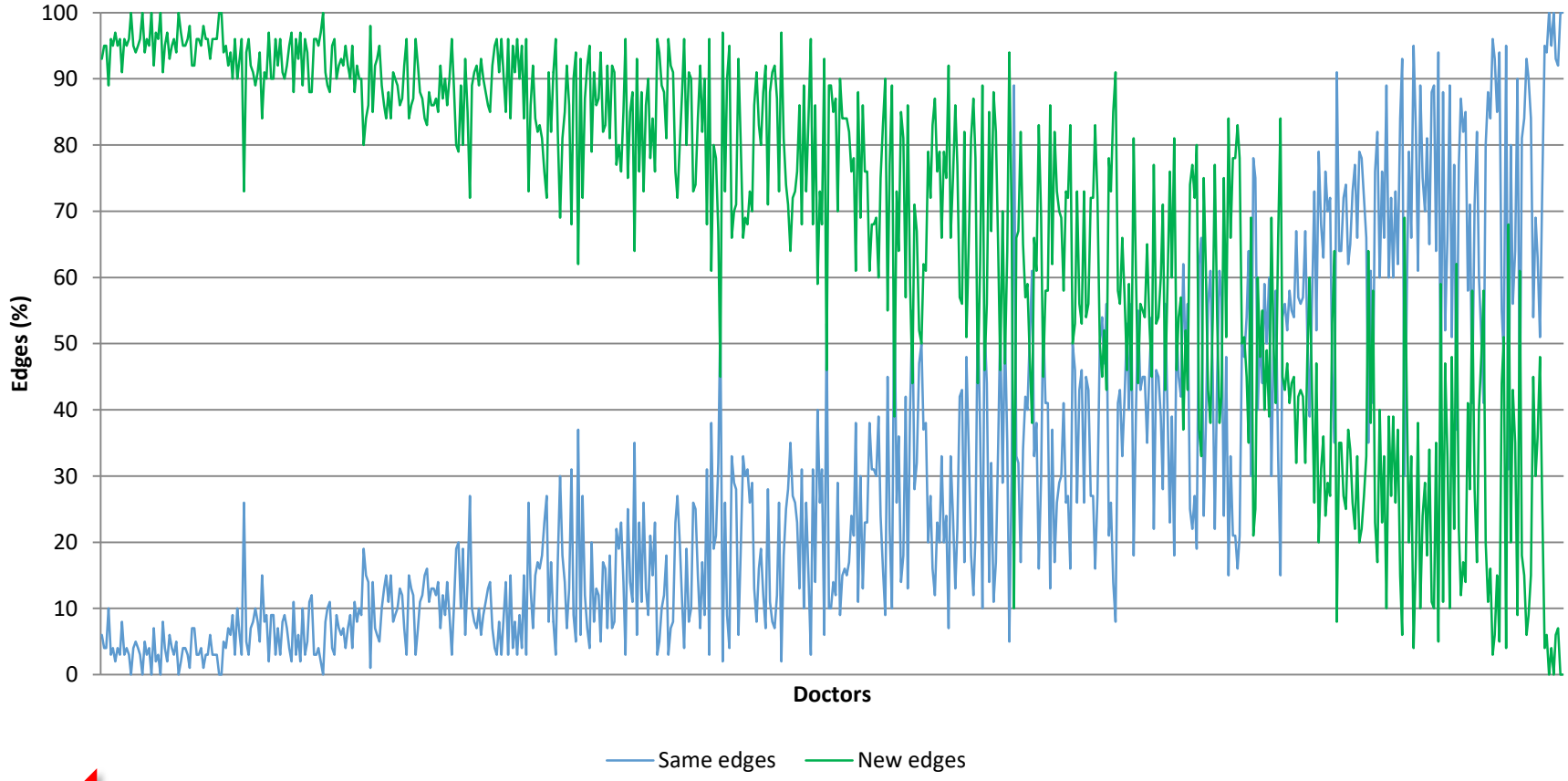


- 23 episodes
- 6 nodes
- 9 edges
- Most frequent paths: <INIT, 4:petEcg, END>, <INIT, 4:petRx, 3:petRx, 5:petRx, END>, <INIT, 4:petEcg, 4:petRx, END>
- Node degree: 3 ± 0,9

- 96 episodes
- 23 nodes
- 65 edges: 7 (same) + 58 (new)
- Most frequent paths: <INIT, 4:petRx, END>, <INIT, 4:petInter, END>, <INIT, 4:petEcg, END>
- Node degree: 5,7 ± 3,2

High distortion
High inf. loss

4. Case Study



Information loss (higher %new edges = new behaviour)

Outline

1. Process Mining
2. Process Mining in Healthcare: Challenges & Opportunities
3. Privacy-Preserving Process Mining
4. Case study
5. Conclusions

5. Conclusions

- Process mining is an emerging field with high potential
- The importance of preserving privacy of individuals in log files
  - Careful management of sensitive information (third parties)
  - Compliance with EU GDPR
- **Future work:**
  - Apply other privacy-preserving techniques (*e.g.* generalization of doctors according to a hierarchy)
  - Exhaustive analysis of much more graph measures (centrality, connectivity, distance…)

Privacy-Preserving Process Mining:

Towards the new European General Data Protection Regulation

Edgar Batista[1] and Agusti Solanas[2]

[1] SIMPPLE, S.L.
C. Joan Maragall 1A
43003 Tarragona, Catalonia, Spain
edgar.batista@simpple.com

[2] Smart Health Research Group
Universitat Rovira i Virgili
Av. Paisos Catalans, 26
43007 Tarragona, Catalonia, Spain
agusti.solanas@urv.cat