STRUCTURING BEHAVIOR OR NOT,
THAT IS THE QUESTION

Wil van der Aalst, RWTH Aachen University, wvdaalst@pads.rwth-aachen.de

Abstract

Process models aim to structure behavior for a variety of reasons: discussion, analysis, improvement, implementation, and automation. Traditionally, process models were obtained through modeling and structure could be enforced, e.g., by streamlining or simplifying processes. However, process discovery techniques that start from the actual behavior shed new light on this. These techniques return process models that are either formal (precisely describing the possible behaviors) or informal (merely a "picture" not allowing for any form of formal reasoning). Both types of model aim to structure reality. However, reality is often very different and much more variable than expected by stakeholders. Process mining often reveals an "inconvenient truth" which provides the valuable insights needed to improve a wide variety of processes. This contribution, devoted to Jörg Becker's 60th birthday, reflects on the notion of "structure" in a world where event data are omnipresent.

Keywords: Process Mining, Business Process Management, Vagueness in Models, Process Discovery.

1   Introduction

It is a great pleasure to contribute to this "Festschrift" devoted to Jörg Becker's 60th birthday. Jörg has been one of Germany's leading "Wirtschaftsinformatiker" for decades and played a key role in the development of the field. He worked on many topics related to information systems (e-business, e-government, information modeling, IT maturity, reference modeling, etc.) and is probably best known for his work on Business Process Management (BPM) (Becker, Rosemann, & Von Uthmann, 2000; Becker, Knackstedt, & Pöppelbuß, 2009; Becker, Beverungen, & Knackstedt, 2010; Röglinger, Pöppelbuß, & Becker, 2012).

Jörg Becker supervised numerous PhD students of which many became very successful in both academia and industry. He created an "IS school" where the credo is: "structure, structure, structure". His guiding principle has been that information system engineering is all about finding a suitable structure. Process modeling and information modeling play a key role in this.

This contribution focuses on the interplay between structure and data (Van der Aalst, 2016). When dealing with real processes, one often finds that process executions follow a Pareto distribution. Some behaviors are highly frequent an easy to capture. However, the "tail of the Pareto distribution" is the real challenge in information system engineering. Although 80% of the process instances may be explained by 20% of the process variants, often most of the resources are put into the remaining 20% of process instances that deviate from the so-called "happy paths".

In the remainder, a simple example is used to show that reality often diverges from simplistic Power-Point models. The makes it far from trivial to structure real-life processes. Process miners typically distinguish between Lasagna and Spaghetti processes. Process models may be viewed as maps that need to be tailored towards specific questions. As such, structuring can be viewed as finding the right map.

2   An Example: Purchase-to-Pay (P2P)

To illustrate the surprising complexity of real-life processes consider the Purchase-to-Pay (P2P) process found in almost any organization. P2P refers to the operational process that covers activities of requesting (requisitioning), purchasing, receiving, paying for and accounting for goods and services. This process is supported by Enterprise Application Software (EAS) from vendors such as SAP, Oracle, Microsoft, and Salesforce. At first glance, this process seems simple, and indeed most cases follow the so-called "happy path" depicted in Figure 1. The activities "create purchase requisition", "create purchase order", "approve purchase order", and "receive order confirmation" are executed in sequence. Then the activities "receive goods" and "receive invoice" can be performed in any order followed by activity "pay invoice" as the final activity.
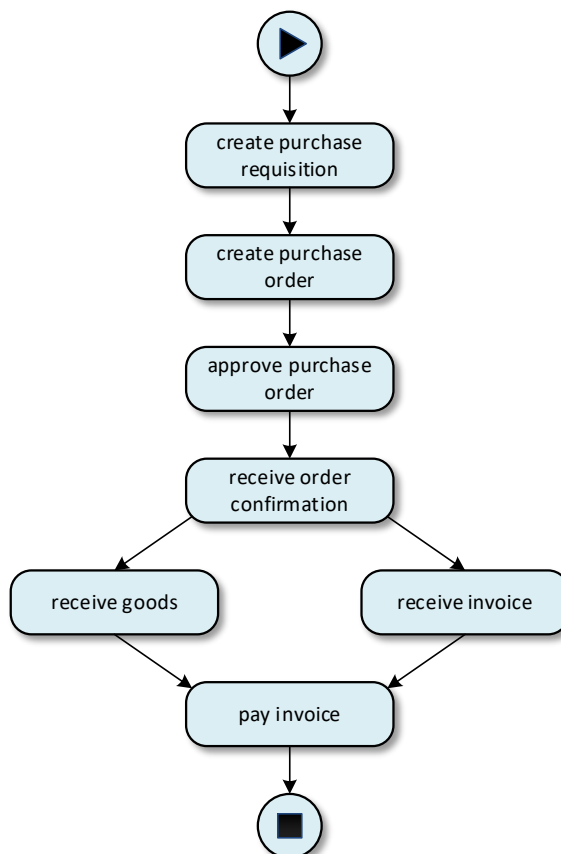
Figure 1: Purchase-to-Pay (P2P) process only considering the "happy path".

The process depicted does not reflect the many variants of the process. Taking a sample of 2654 cases (i.e., purchase orders) and showing all the paths reveals the true complexity of the process. Figure 2 shows the so-called directly follows relation. Here we can see which activities follow one another. The 2654 purchase orders follow 685 unique paths. Clearly, the cases follow a Pareto distribution. The most frequent path is taken by 201 cases. The second most frequent path is taken by 170 cases. 68% of the variants are unique and account for only 17% of the cases. 63% of the cases can be explained by 8% of the variants, and 82% of the cases can be explained by 31% of the variants. Hence, the distribution approximates the well-known 80-20 distribution. Note that this example is not exceptional. This holds for most P2P processes and also applies to similar processes that are not fully controlled by software.

Process mining techniques can cope with such complexities (Van der Aalst, 2016). By removing some of the infrequent paths, we can find the process model depicted in Figure 3. Such a model can also be

translated to a Petri net, BPMN model, UML activity model, or EPC. The model can be further simplified setting thresholds on frequencies.

The different process variants may have very different behaviors, not only in terms of control-flow, but also in terms of Key Performance Indicators (KPIs). For example, a price change may add a delay of 4.5 days on average. Infrequent paths may point to fraud. For example, orders that were paid but never delivered.
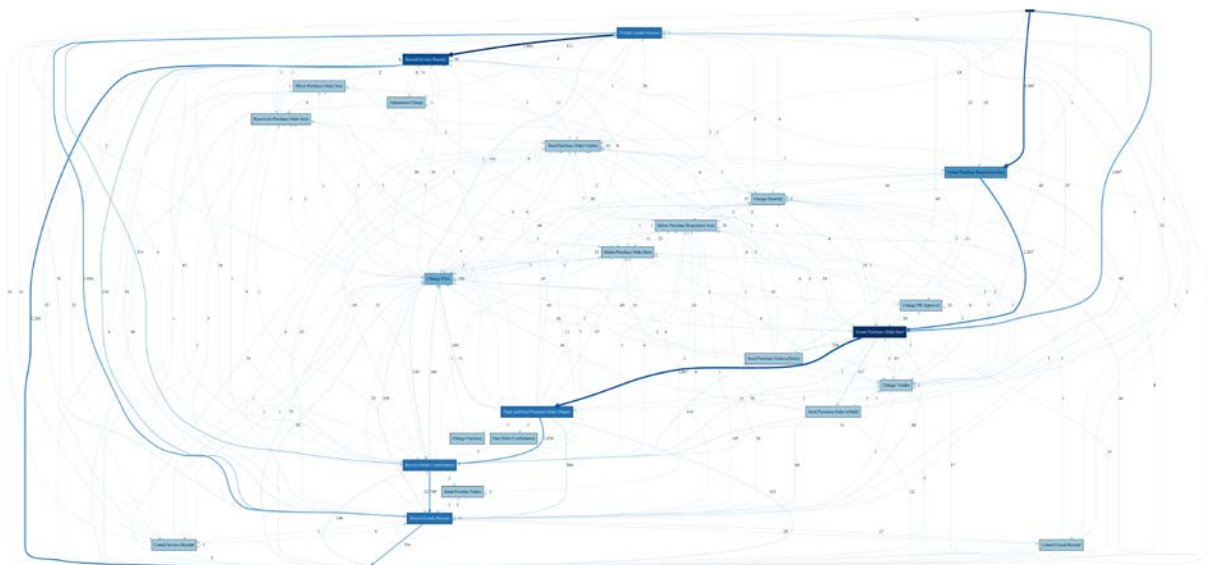

Figure 2: The real P2P process: 2654 purchase orders follow 685 unique paths.
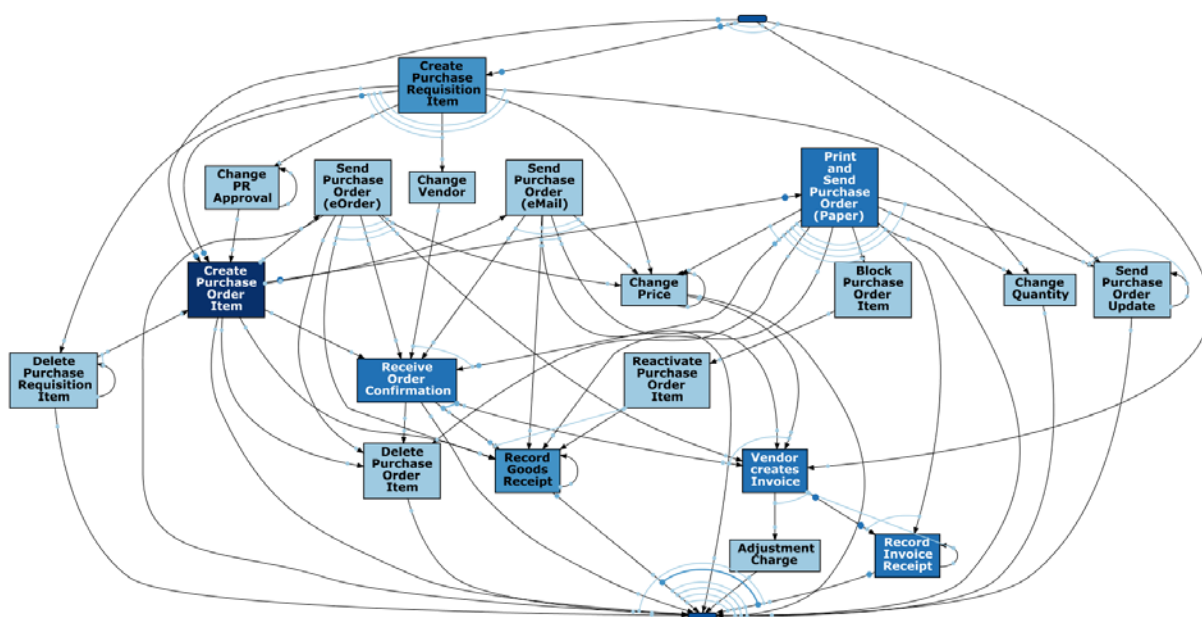

Figure 3: A so-called Causal Net (C-Net) describing the process model.

3 Between Lasagna and Spaghetti

The simple P2P process shows that reality may be surprisingly different from reference models and PowerPoint diagrams. The terms Lasagna and Spaghetti refer to the different types of processes. A simple metric is the number of process variants (unique traces) divided by the number of cases. This

yields a number between zero and one. The closer to one, the more Spaghetti-like the process is. The closer to zero, the more Lasagna-like the process is. For the P2P process discussed, the metric is 685/2654=0.2581. This is one of many ways to characterize event logs and the underlying processes.
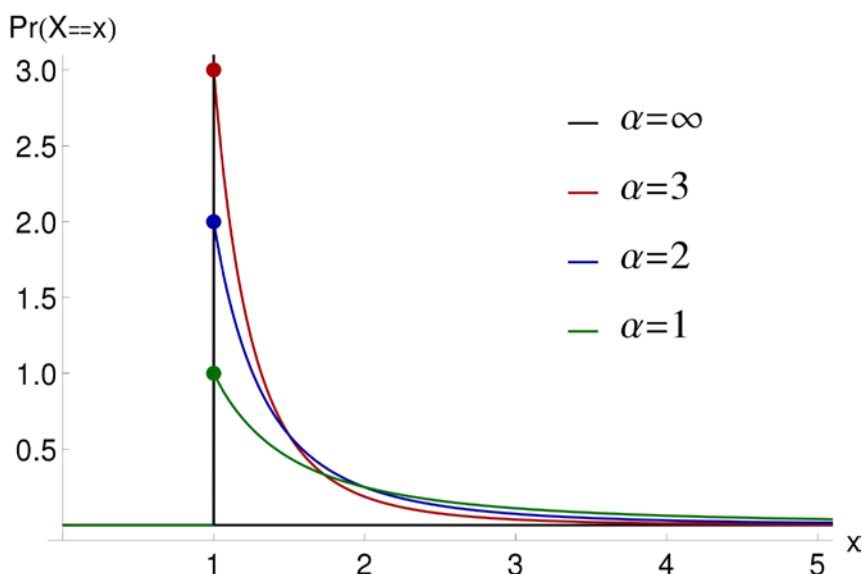
Figure 4: Pareto Type I probability density functions for various α values.

Figure 4 shows the Pareto Type I probability density function for various values of α. The x-axis corresponds to the different traces (unique behaviors) sorted by frequency. The y-axis represents the relative frequency of each trace. The higher the value of α, the more uneven the distribution. Note that the distribution has a "head" (left-hand part of the distribution composed of the most frequent cases) and a "tail" (right-hand part of the distribution composed of the less frequent cases). The tail is often long. Analysis may focus on the head (e.g., when improving performance) or the tail (e.g., when dealing with compliance problems). This shows that the boundary between Lasagna and Spaghetti is not so clear-cut. Even within the same process, one can find both types of behaviors.

4   Structuring = Finding a Suitable Map

So how does this relate to Jörg's credo "structure, structure, structure"? It is not so easy to find structure when dealing with real-life processes. However, it remains important to look at the problem from the right angle. One can view process models as geographic "maps" describing reality. A subway map looks very different from a bicycle map although they aim to describe the same city. What is the best map? This depends on the purpose. The same holds for process models. What is a good model? This depends on the questions it intends to answer. The large availability of event data allows us to seamlessly generate and use process models in ways we could not imagine in the 1990s. However, the challenge remains to find structure.

Process discovery techniques that start from the actual behavior shed new light on the suitability of process model notations. There is a gap between techniques that return formal process models (precisely describing the possible behaviors) and techniques that return imprecise process models ("pictures" not allowing for any form of formal reasoning). However, parts of a process may be clearly structured, whereas other parts are not. Hybrid process models have formal and informal elements, thereby exploiting deliberate vagueness (Van der Aalst et al, 2017). One should not try to structure behaviors that have no structure; otherwise, one there is the risk of overfitting the data. Applications of process mining clearly demonstrate the advantages of being precise when possible and remaining "vague" when there is not enough "evidence" in the data or standard modeling constructs do not "fit"

(Van der Aalst et al, 2017). We envision that the next generation of commercial process mining tools will support such hybrid models.

To conclude, I would like to congratulate Jörg again with his 60th birthday! A milestone in a remarkable career.

References

Van der Aalst, W (2016). Process Mining - Data Science in Action. Springer Verlag.

Van der Aalst, W., De Masellis, R., Di Francescomarino, C., & Ghidini, C. (2017). Learning Hybrid Process Models from Events - Process Discovery without Faking Confidence. BPM 2017: 59-76.

Becker, J., Rosemann M., & Von Uthmann C. (2000). Guidelines of Business Process Modeling. Business Process Management 2000: 30-49.

Becker, J., Knackstedt, R., & Pöppelbuß, J. (2009) Developing Maturity Models for IT Management. Business & Information Systems Engineering 1(3): 213-222.

Becker, J., Beverungen, D., & Knackstedt, R. (2010). The challenge of conceptual modeling for product-service systems: status-quo and perspectives for reference models and modeling languages. Inf. Syst. E-Business Management 8(1): 33-66.

Röglinger, M., Pöppelbuß, J., & Becker, J. (2012). Maturity models in business process management. Business Process Management Journal 18(2): 328-346.