Operational Research (2019) 19:887–908
https://doi.org/10.1007/s12351-018-00447-z

ORIGINAL PAPER

Visualizing and exploring event databases: a methodology to benefit from process analytics

Pavlos Delias[1] · Vassilios Zoumpoulidis[1] · Ioannis Kazanidis[1]

Received: 27 April 2018 / Revised: 3 October 2018 / Accepted: 17 December 2018 /
Published online: 10 January 2019

Abstract
Events, routinely broadcasted by news media all over the world, are captured and get recorded to event databases in standardized formats. This wealth of information can be aggregated and get visualized with several ways, to result in alluring illustrations. However, existing aggregation techniques tend to consider that events are fragmentary, or that they are part of a strictly sequential chain. Nevertheless, events' occurrences may appear with varying structures (i.e., others than sequence), reflecting elements of a larger, implicit process. In this work, we propose a methodology that will support analysts to get richer insights from event datasets by enabling a process perspective. Through a case study about a political phenomenon, we provide concrete recommendations on data reviewing, process discovery, and visually facilitated interpretations. We furthermore discuss the methodological and epistemological aspects that are needed to make our approach applicable for event analytics.

Keywords Event data · Process mining · Process analytics

1 Introduction

Discovering knowledge through event databases is a challenge that is being pursued for decades. The first motivation seems to originate from politics, since in the 60s, McClelland (1961) tried to analyze political interactions at a macro-level. However, several years were needed for this initiative to reach an accomplished version, the codebook of WEIS (McClelland 1976). The WEIS was perhaps the first systematic approach to structure event data, which allowed a departure from the idiographic study of political events, towards the use of quantitative methods. The importance of having a systematic method to listen and to record political events, was that big, that different initiatives (including projects funded by Defense Advanced Research Projects Agency, and National Science Foundation) produced systems like KEDS

✉ Pavlos Delias
pdelias@teiemt.gr

[1] Eastern Macedonia and Thrace Institute of Technology, Kavala, Greece

(Gerner et al. 1994), CAMEO (Gerner et al. 2002), ICEWS (O'Brien 2010), and recently GDELT (Leetaru and Schrodt 2013). The long sought out goal of all these efforts is the discovery of regularities of events occurrences.

The aim of this work is to enrich the toolbox of analysts that look for this kind of regularities by enabling a process perspective for event data. Aiming at delivering rich descriptions, we provide a guide on reviewing event datasets, empowering knowledge discovery techniques, and facilitating the analysis with relevant visualizations. We shall notice of course that by "process perspective" we refer to the structural aspects, since elements of the functional context of business process management (e.g., redesign, execution) are not applicable to event analytics (at least to social/political event analytics).

Discovering knowledge through event databases becomes more and more complex, and interesting, since the primary obstacle in acquiring such data (manual collection and coding procedures) has been mainly solved. Currently, virtually all news articles that can be spotted by web crawlers are machine-coded, and are getting registered. Therefore, the challenge has shifted to the analysis side. Lately, many event analytics techniques with appealing visualizations have been developed (see Sect. 2 for a brief overview). However, these (rich) techniques seem to univocally focus on the pure sequentiality of events occurrences. The dominant ways to portray regularities is Frequent Sequence (or Pattern) Mining (applied to sequences of temporally ordered events) (Gotz et al. 2014), and flow diagrams (Liu et al. 2017), like Sankey plots, which in the case of small data (i.e., few events per case and limited event alphabet) are producing alluring illustrations.

Nevertheless, real-world news events often do not follow a linear movement that is implied by a sequence, but they have varying structures, and these may reflect elements of a larger process that has many descriptive components (Peuquet et al. 2015). In this work, the ambition is to enable the observation of the events through a process perspective. We advocate that such an activation will allow analysts to consider questions common in process diagnostics (Bose and van der Aalst 2012) for event analysis. In particular, we expect to be able to get answers to questions like: (1) Are there any paths, of any size, that reveal some kind of frequent interaction patterns? (2) Can we manifest decision points, or deviating behaviors by introducing gateway semantics (flow splitting and joining points) on events' occurrences? (3) Does the thematic characterization of events alter the interaction patterns? (4) Do the involved actors interact based on any patterns? (5) Can we recognize any properties for the events (or actors), e.g., milestone events, hub or authority events or actors?

The key components of our approach are quire straightforward: Since event data standards qualify similar formats (that include a source actor, a target actor, an event, and a timestamp), it is clear that any approach that has a slightest aspiration to make an impact, must exploit this standardized format. Therefore, we propose transforming the original format into various options, all of them enabling event data formats to get loaded into *Process Mining* (van der Aalst 2016) tools. Process mining techniques will eventually allow process models discovery. These (automatically discovered) process models can be effectively illustrated, and conceivably deliver the contributions we claimed in the previous paragraph. In addition, a plethora of decision support methods (already proposed for original process mining applications) will be able to get leveraged for event analytics.

We should note that our approach does not intend to deliver a theory or a general model but rather a methodology and a set of recommendations on how to get richer insights for event analysis. The epistemological considerations of such an approach are discussed in Sect. 3, in tandem with the methodological approach, and the presentation of the case study that we are going to use to illustrate our proposal. In Sect. 2 we provide a brief review of existing approaches for event analysis, as well as for relevant process mining methods. Our proposal about how a process orientation can be infused into raw event datasets is presented in Sect. 4, while a short discussion concludes the paper.

2 Background

The “Global Data on Events, Location, and Tone” (GDELT) project, supported by Google, consists of hundreds of millions event records, extracted from broadcast, print, and online news sources from all over the world (Leetaru and Schrodt 2013). The GDELT provides a free, rich, comprehensible, daily updated, CAMEO-coded (Gerner et al. 2002) dataset, which applies the TABARI system (Best et al. 2013) to each article to extract all the events disclosed therein.

This valuable resource has been exploited in many research efforts (e.g., Keertipati et al. 2014; Kwak and An 2016; Ward et al. 2013; Phua et al. 2014), however, event data are treated either as input to ordinary analytics techniques, or as time-series data (Jiang and Mai 2014). Considering event analysis, the most visible methods are survival analysis and event history analysis (Broström 2012). Classical survival analysis focuses on research questions involving the time elapsed from a start event to an event of interest. This is an interesting research problem because of censoring and truncation (deriving form incomplete observations of events). To calculate the survival function and hazard rate (the most popular outcomes of research questions) several methods like the Kaplan–Meier and the Nelson–Aalen estimators, regression models (e.g., Cox and Poisson), parametric models (e.g., proportional hazard or accelerated failure time models), or even frailty models have been developed. However, it is often the case that events of interest occur more than once for an individual (recurrent events), or there are multiple types of events of interest. Event history models arise to allow following subjects over time and making notes about what happens and when (Aalen et al. 2008).

Indeed, treating event data like sequences, can yield effective visualizations that could support decision makers. Several methods for querying, filtering, and clustering multiple event sequences have been proposed, for example Fails et al. (2006), Vrotsou et al. (2009) and Wongsuphasawat et al. (2012), or the works of Gotz and Wongsuphasawat (2012) and Wongsuphasawat and Gotz (2012) that can handle much larger numbers of event sequences and provide effective visualizations for their aggregations. Moreover, when these methods can be combined in a user-friendly dashboard, decision support can be further improved (Gotz and Stavropoulos 2014). For a concise description of how event information can be modeled, retrieved, and analyzed, the reader is directed to Gupta and Jain (2011).

Nevertheless, if we assume a process perspective (i.e., that events are not happening in random, but their occurrence is part of a larger, implicit process), the *process mining* paradigm is enabled, and it substantially augments the decision support potentials. In particular, the following competences will be facilitated:

- Discover complex structures of events (splitting and merging points, long-distance and multi-step causalities, etc.) even when the process is drifting over time (e.g., due to recurring seasonal effects) (Martjushev et al. 2015)
- A family of process mining techniques aims at detecting and explaining differences between executions that lead to different outcomes. Under the general term deviance mining (Nguyen et al. 2014), we can see for instance, approaches that return in natural language the significant differences between traces that lead to a special outcome and traces that don't (van Beest et al. 2015), or point out the factors that differentiate the flows (De Leoni et al. 2014; Delias et al. 2015b). Similar behavior (which eventually leads to similar results) can be also identified through trace clustering (Song et al. 2008; Bose and van der Aalst 2009), where trace profiles are created based on control-flow, organizational, performance or additional criteria, and then traces are grouped according to some similarity metric. Trace clustering techniques are particularly useful to unclutter process models when a lot of variation exists.
- Check deviation from expected pathways. Another type of process mining is conformance checking, where an existing (ideal) process model is compared to the event log of the same process. Conformance checking can be used to check if reality, as recorded in the log, conforms to the model and vice versa (van der Aalst et al. 2012). Local or global diagnostics can be created to check deviations form the expected pathways, or even to check the effect that possible "history modifications" would have to the discovered model (van der Aalst et al. 2015).
- Putting timestamped events into a process structure allows to observe the temporal evolution (performance) of the process. In process mining, this family of methods is known as performance analysis (van der Aalst et al. 2011; Adriansyah and Buijs 2012), and can respond to questions like: how process performance evolves? Are some events delayed due to bottlenecks? Would the resolution of bottlenecks or following some special paths accelerate some events (Nguyen et al. 2016)?

With respect to our knowledge, this is the first time that a framework that enables a process perspective on event data is proposed. This is a significant contribution to the field, since a whole novel family of methods will be enabled.

3 Preamble of the approach

3.1 Case study

July of 2015 was no ordinary month for Greece. Following a dubious negotiation strategy, the Prime Minister Alexis Tsipras, on the 27th of June 2015, announced a referendum to decide whether Greece was to accept the bailout conditions, proposed by the European Commission, the European Central Bank, and the International Monetary Fund. Although the result was a clear "No" (61.3%), one week later, Greece's parliament signed on to harsher bailout terms, i.e., a new mid-term Memorandum of Understanding. News was coming rapid and spectacular, since the government soon called an early parliamentary election. The story has been extensively covered by international media, making it eminently suitable for an illustrative example for the proposed approach. Dozens of thousands of events were registered to the GDELT, allowing for rich and informative event analytics.

Several techniques and visualizations are provided through the GDELT Analysis Service (http://analysis.gdeltproject.org). For example, by exploiting the geographical tagging of events, it is possible to plot a heat-map which will help us to understand the spatial patterns of the events. Moreover, the GDELT constructs the so-called Global Knowledge Graph (GKG) by extracting information such as the connections of persons, organizations, locations, emotions, and themes identified in the source texts, making it possible to generate word clouds of the most popular themes identified in the events, or tone (emotional connotation of the words in the article) timelines. As authors in Delias and Kazanidis (2017) illustrate, in that period's broad-casted news about Greece, the term "tax" dominates, while other terms that stand out are "debt", "bankruptcy", "negotiation", "fragility", all of them accurately reflecting the situation of that period. In addition, the announcements for the referendum and the early elections, caused two sudden drops of the tone (the negative emotions become even more negative).

All these interesting visualizations provide interesting insights, derived from various aggregation perspectives of the events. However, in all of them, events are considered fragmentary, "rambling" elements of the story. Should anyone develop a hypothesis that events occur as part of a regularity, i.e., a process, these kind of visualizations can not support the relevant checking. In the following sections, we present our proposal for a methodology that responds to this challenge, and allows analysts to exploit the richness and availability of raw event data. Ultimately, the proposed methodology enables a process perspective for the (otherwise disjointed) events.

3.2 Methodology and epistemological considerations

We should note that in our approach, we make the fundamental assumption that the notion of a process is relevant for the realizations of events, i.e., that there is some kind of rational structure over the events, and therefore the challenge is to unveil this structure. Therefore, our approach follows the prescriptive paradigm that suggests discovering a model, suitable for a given situation in a particular context, and does not intend to be general (like, for instance, in a descriptive approach, which would have aimed at deriving global laws from the observed phenomena). Therefore, should we try to position epistemologically our work, we could say that it is located between the bottom-line which aims to produce a narrative based on empirical observations (namely rich descriptions of the phenomenon), and the ambition to develop a model with some level of abstraction. This area shares a portfolio with the grounded theory methodology, at least with its implementations in information systems research (Wiesche et al. 2017), yet we do not claim that our work can be classified as a grounded theory methodology approach mainly because of the big differences in the data collection and analysis procedures.

Considering the methodological aspects, we build on the decision support paradigm (Roy 1994), and we follow a rather more contextual and operational approach, inspired from Delias et al. (2015a). More concretely, in Delias et al. (2015a), authors, following the process mining paradigm, propose a methodology based on a practical synthesis of the common methodological steps of existing approaches to guide the implementation of process analytics projects. In particular, the original methodology suggests four basic phases:

1. *Business understanding*, including actions like defining project's scope, uncovering facts, constraints, and assumptions, aiming at figuring out the business context and developing the shape of solutions
2. *Data collection and reviewing*, where actions like filtering, imputing missing data, data transformations are performed
3. *Discovery*, which includes a series of techniques intending to extracting knowledge relevant to the project's objectives
4. *Decision aid*, in terms of actions like proving recommendations, deploying the solutions, etc. meaning to build a rapport between results and business goals.

In this work, we endorse that methodology, by presenting a refinement. More specifically, we exemplify actions of the first two phases by adjusting them into the event databases context. However, the prime focus of our refinement lies in the discovery phase, wherein we discuss what are the particularities of analyzing event databases with process mining techniques, and we reveal the benefits of such a resolution. From a procedural point of view, since the focus is on communicating insights from the data analysis, we could label our approach as "memoing" (Glaser 1978), conditional of course on the research goals, which are to deliver rich descriptions and not a theory. Finally, we leave out of the discussion the last phase, the decision aid, since the political nature of the events that are involved, requires for high-level governmental conduct, making the decision aid phase out of scope for this paper. A figurative register of the steps that this methodology comprises to deal with the analysis of case study is illustrated in Workflow 1. The next section instantiates this workflow and describes the details of its application.

Workflow 1: Getting insights by applying a process perspective

```
Stage 1: Business Understanding
    Define project objectives ;
    Define project scope;
    Define process scope;
Stage 2: Data Collection and Reviewing
    Query sources to get raw data $\mathcal{R}$;
    Find rows with empty actors $\mathcal{R}_{EA}$ : $\mathcal{R}' \leftarrow \mathcal{R} \setminus \mathcal{R}_{EA}$ ;
    Remove duplicates from $\mathcal{R}'$: $\forall r_i, r_j \in \mathcal{R}', r_i \neq r_j$;
    Find rows with just 1 activity $\mathcal{R}'_{1A}$ : $\mathcal{R}'' \leftarrow \mathcal{R}' \setminus \mathcal{R}'_{1A}$ ;
    Remove irrelevant rows (noise) from $\mathcal{R}''$ ;
    Event Log 1 $\mathcal{EL}_1 \leftarrow$ Transform( $\mathcal{R}''$, $template1$);
    Event Log 6 $\mathcal{EL}_6 \leftarrow$ Transform( $\mathcal{R}''$, $template6$);
return $\mathcal{EL}_1$,$\mathcal{EL}_6$
Stage 3: Discovery
    ProcessMap($\mathcal{EL}_1$,args:Activities coloring);
    ProcessMap($\mathcal{EL}_6$,args:Theme filter);
    Trace Clustering (empirical control of similarity metrics);
    Discover Patterns instead of models ;
return memoing
```

4 Getting insights by applying a process perspective

4.1 Business understanding

Although the term "business" seems a bit unorthodox to describe political events about countries interactions, we shall preserve it to address the general situation of describing the context of the problem. In that sense, describing business objectives takes the shape of specifying the expected benefits in the context terms. Therefore, in this work, the objective is to reveal regularities in the occurrences of political events among countries, as an additional tool to perform more refined empirical analyses of political and/or social operations. The case study limits the scope on a specific time period (the period relevant to the Greek referendum of 2015), and to specific countries (the countries relevant to the involved institutions).

However, besides defining the project's scope, an additional scoping activity, which is recommended to be part of the "understanding" phase, and which is compatible with our process-oriented perspective, is defining the *process* scope. According to van der Heijden (2012), defining process scope deals with acknowledging what parts of the process can be tracked through the logged data and what type of information is both available and useful. In this regard, it is of primary importance to consider the original event data format.

The standard format of raw event data (e.g., ICEWS, GDELT) comprises at least the following basic elements: A timestamp, two fields indicating the involved actors (the one treated as the source, and the second as the target), and a code for the pertinent action that took place. For example, Table 1 shows a snippet of a GDELT record, that happened on the 2nd of July 2015, and involved Greece (Actor code: GRC) and refugees from Syria (Actor Code: SYRREF). The event that took place is the 043: "Host or receive a visitor at residence, office or home country", but it doesn't refer to the actual arrivals of refugees (it would be quite ironic to call that a "visit"), but to actors from the popular TV show "Game of Thrones" that visited Greece to Call on EU Leaders to Help Refugees Stranded in Greece (the original news can be found at: http://people.com/tv/game-of-thrones-stars-call-on-eu-leaders-to-help-refugees-stranded-in-greece/). Many additional fields (e.g., the coordinates for the location of the event, its tone, the url of the article) are recorded as well. However, this format can not be directly matched to a format that will enable a process perspective, such as the input requirements of *Process Mining*. In order to make it useful for our objective, the activities of the following phases are proposed. In any respect, we shall recall from Sect. 3.2 that a fundamental assumption of this work is that there is some kind of rational structure over the events' occurrences, and that this structure can be observed through the event datasets.

Last, a meaningful activity for the "understanding" phase is to suggest a set of performance indicators, which ideally will match indicators to objectives, and which provide a basis of comparison for the delivered results. Since following this methodology we expect to expose regularities of events' occurrence, metrics such as support, confidence, language fit, determinism, and coverage, proposed by Tax et al. (2016a) might become relevant. However, following the "memoing" procedure that we suggested in Sect. 3.2, the need for quantitative indicators can be relaxed. We can therefore accept as a legitimate result whatever contributes in building a rich description of the phenomenon.

4.2 Data collection and reviewing

To collect a relevant dataset, we queried the GDELT database. We performed three queries: one that returned all events that happened between the 20th of June 2015 and the 20th of July 2015 (the contested period of the referendum) and whose source actor's country was Greece. The second query was similar, with the single difference that we asked for the *target* actor's country to be Greece. The third query asked for all events that happened during the same period in Belgium (the host country of E.U. institutions, and where most official negotiations took place), without specifying any actors' countries. These queries returned a dataset of 30,000 rows. Apparently, this dataset is not complete (it will take far more countries and longer time period to reach a more complete dataset), yet it is a sample that can sufficiently demonstrate our approach.

An integral activity of the *data collection and reviewing* phase is *filtering*. Although the filters to be applied on event datasets will always be case-dependent, we can recommend the following ones:

Table 1 A sample record of GDELT (only basic fields are showing)

SQLDATE	Actor1 code	Actor2 code	Event code
20150702	GRC	SYRREF	043

- Since actors (either source or target) are an indispensable variable of the process perspective (see Sect. 4.2.1 below), records that do not include any information about the involved actors can be removed. In our case, because we care about the interactions at a country level, we removed the records that were blank in either the source or the target (or in both) actor's country field.
- It is highly possible that different media had captured the same event, so it is equally possible for the dataset to contain many duplicates. Such duplicated rows were removed.
- Embracing a process perspective signifies that we care on sequences of events. Therefore, if a case (see Sect. 4.2.1 about how a case can be defined) consisted of just one event, it was not not helpful for our process-oriented analysis, and hence it was removed.

In addition, we manually browsed the url of the events, to detect irrelevant ones. Actually, we encountered several extraneous events that were removed, yet some noise remained. Popular themes of the extraneous events were Brexit, refugees migration, terrorist attacks, or even recreational events like football players' movements, and concerts.

Finally, since the focus of this illustrative example is on countries interactions, we chose to remove the events that included the same country both as source and target actor, a pattern that commonly indicates the debates between opposition and government within that country, and clutters the countries interactions' visualizations. Applying all the above filters resulted in a dataset of 4476 events, which indicates the necessity of filtering as a data reviewing activity.

4.2.1 Transformation of event data

As we have discussed in Sect. 4.1, the standard format of event data can not be directly matched to a format that will enable a process perspective. More specifically, Process Mining requires at least three basic fields for every record: a *case ID* (to correlate events with a particular case), a *timestamp*, and an *activity* (van der Aalst 2016). All applications of Process Mining on different types of data, from low-level machine or software logs (Günther et al. 2009; Mannhardt et al. 2016) to incidents' status changes in CRM systems (van Dongen et al. 2013), assume this data format. This work proposes the following mappings to enable event data to get exploited by process mining tools:

As long as it concerns the timestamp field, the pairing is clear: it can be directly matched to the corresponding field of the raw event data. However, it is not clear at all what will be the *case ID*, and/or the *activity*. In Table 2, we provide six alternative mappings that can be applied, each of them delivers a different view of the data, yet all of them administer a process-oriented view.

The first alternative transformation uses the combinations of the two actors (hence the underscore as the joining delimiter) to distinguish the case identifiers, and the event code field to create the alphabet of activities. This way, a case is a

Table 2 Alternative transformations to match raw event data to process mining input requirements

	Case ID	Activity
1	SourceActor_TargetActor	Event code
2	SourceActor	Event code
3	TargetActor	Event code
4	Event code	SourceActor
5	Event code	TargetActor
6	Event code	SourceActor_TargetActor

bilateral relationship of two actors, which leaves a trace of events that happened and involved both of them. By transforming raw event data in this format, we can use the dataset as an input for automated process discovery.

The next two proposed alternative transformations are similar, in the sense that they still make use of the event code as the activity, and an actor as the case ID, yet these transformations use a single actor (i.e., not a combination). These alternatives are suitable for datasets where many countries participate, and there is a limited thematic selection. In such situations, we expect process maps to reveal if there are any behavioral patterns for single countries (e.g., if countries that "criticize or denounce" are "using conventional military force" as a following action).

Transformations 4, 5, and 6 inverse the logic. Traces are now joined by the event code (case ID). The items that each trace comprise (activity) are either single actors (source or target) or their combinations. This inverse logic allows to observe the interactions of actors subject to a special thematic. In Sect. 4.3 we provide illustrative instantiations for the relevant templates.

4.3 Discovery

Following the first transformational template of Table 2, and with the support of Celonis Academic Cloud (Celonis 2017), we created the illustration of Fig. 1. We shall note that only the most frequent activities and transitions are illustrated. Since, as the first template suggests, every case is a pair of countries, the process control-flow map actually plots regularities of events' occurrences, responding this way to the first question we mentioned in the Sect. 1 ("Are there any paths, of any size, that reveal some kind of frequent interaction patterns?") . Moreover, we have colored the activities according to their CAMEO code category (see the legend of Fig. 1).

We shall point out two interesting patterns: First, the "Consult" activities are connected either to other "Consult" activities or they are interleaved to "Make public statement" activities. Since statements are typically subordinate events (events are assigned with a code of that category only when they do not further imply appeals, agreements, support, apologies, demands, disapprovals, rejections, threats, etc.) and because "Consult [NS]" is assigned when the place of the meeting is not explicit in the lead, (so no visit made or hosted can be extracted), and

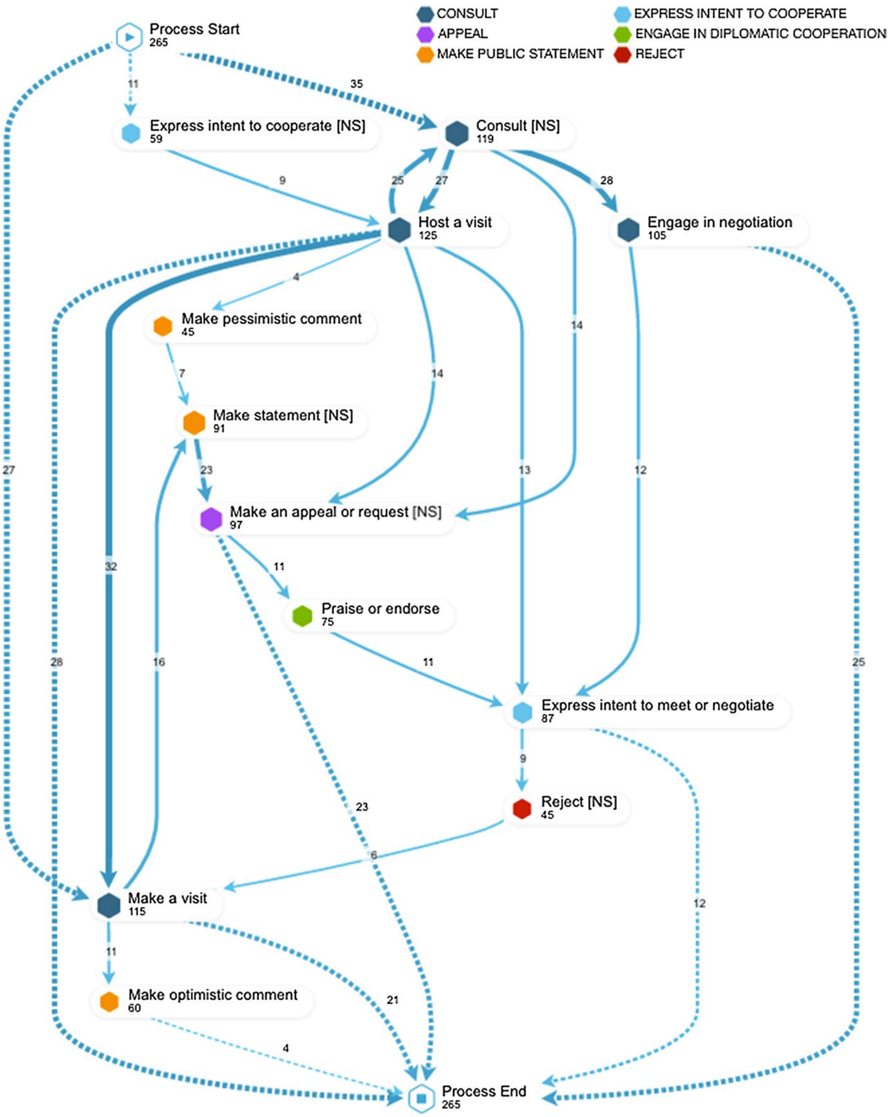

Fig. 1 Process map following the transformation template 1 of Table 2. The color of the event nodes is after their thematic characterization according to CAMEO. Only the top-12 most frequent events are showing. The visualization was created with Celonis Academic Cloud (Celonis 2017) (color figure online)

when no negotiations are implied, we can consider this pattern as a preparatory phase of the institutional discussions.

Next, hosting a visit, when not leading to another “Consult” activity leads to either making a pessimistic comment, or to making an appeal or request, or to expressing intent to meet or negotiate, a pattern that competently reflects the turbulent political situation of that time when hosting a visit did not immediately mitigate the political tension that existed the days before or after the referendum.

Moreover, by visually examining Fig. 1, we can observe that the activity "Host a visit" acts like a *split* gateway since the flow at that point diverges (there are 6 outgoing flows from it). The inverse (several incoming flows) can be observed at the "Express intent to meet or negotiate" activity, which acts like a *join* gateway. This kind of visual interpretation offers a quick response to the second question we described in the Sect. 1 ("Can we manifest decision points, or deviating behaviors be introducing gateway semantics (flow splitting and joining points) on events occurrences?")

To capture the interactions of actors subject to a special thematic, we may use the transformations 4, 5, and 6 of Table 2. The thematic characterization is the focus of the third question we mentioned in the Sect. 1, so we exemplify the potentials of our methodology with the sixth transformation, which is of particular interest for our illustrative example. We demonstrate this kind of potentials with Fig. 2. Figure 2 presents what pairs of actors were involved in "Make Public Statement" events (e.g., make pessimistic/optimistic comment, acknowledge or claim/deny responsibility, make empathetic comment). An interesting pattern that emerges is the path that connects the pair of EU and Greece, followed by Greece and USA, followed by Greece and EU, eventually followed by Germany and Greece. This path either acts like a closed loop, or it escapes to other active EU member states (e.g., France, Italy). This path reflects a common situation of that time: European authorities making a statement about their opinions/decisions on Greece, then Greece making a comment about the IMF, and then watch the member states (the protagonists) making refined commentaries.

Another advantage of plotting the events with a process-oriented perspective is that its representation via a network-like structure can lead us to the recognition of properties for them (like it is expressed in the fifth question in the Sect. 1). More specifically, it is possible to calculate the hub or the authority score (centrality) for any activity. Authority is expected to assess the importance of the activity, whereas hub is expected to reflect the value of its links to other activities. An activity has a high authority score if it is pointed by many activities with high hub scores whereas an activity has a high hub score if it has pointed to many activities with high authority scores. In that respect, Tables 3 and 4 reveal additional insights. For example, in Table 4, we see that the node with the highest authority is the node "USA-GRC" which stands for the interactions of the IMF with Greece. It is worth mentioning that this node while it has the top authority for the "CONSULT" thematic, it has a lower authority when the thematic is about "MAKE PUBLIC STATEMENT" (see Fig. 2), where the top authorities are the nodes "DEU-GRC" and "BEL-GRC". This is an indication that when considering "consult" activities, the IMF was the protagonist, but in making statements, that role was played by the European partners. This is of course an additional hint how thematic characterization influences the interaction patterns (recall question (3) in the Sect. 1).

The control-flow maps, like the ones illustrated in Figs. 1 and 2, can render events' occurrences into meaningful structures, and designate regularities, however,

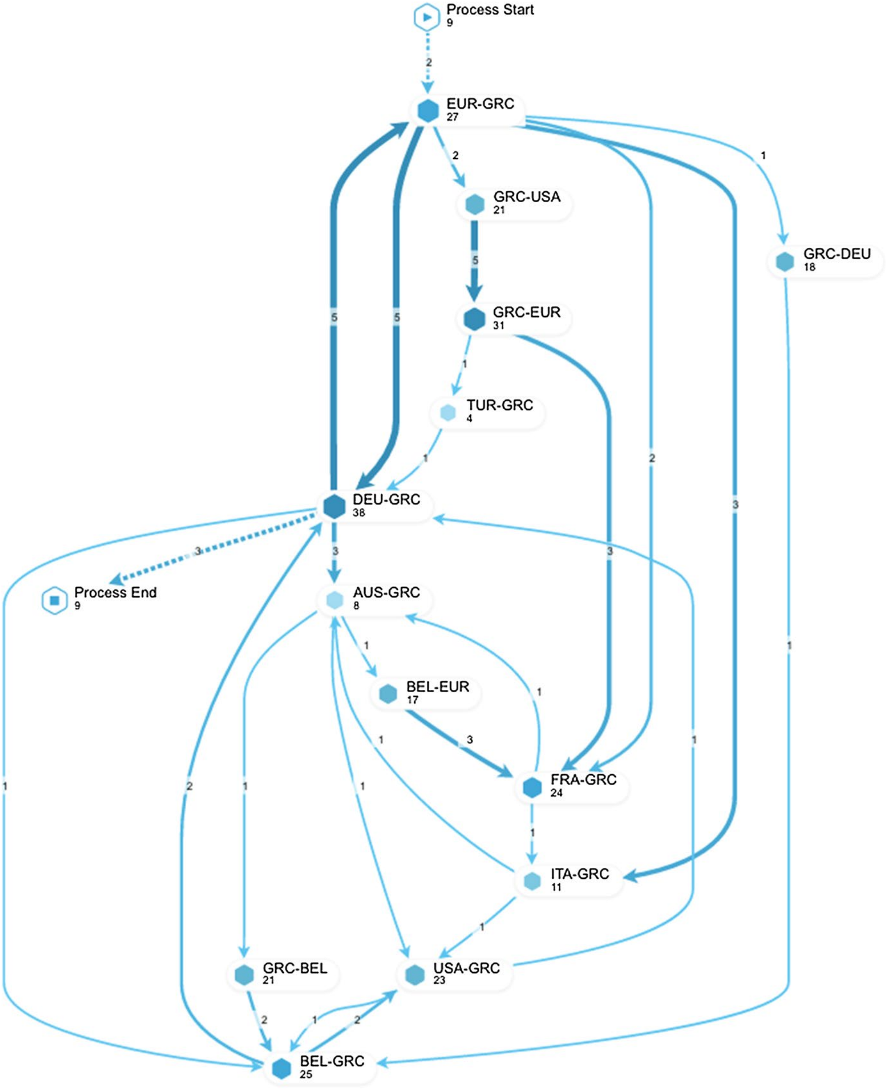

Fig. 2 Process map following the transformation template 6 of Table 2. The cases are filtered to the "MAKE PUBLIC STATEMENT" thematic category. Only the top-12 most frequent countries' pairs are showing. The visualization was created with Celonis Academic Cloud (Celonis 2017)

Table 3 Activities with high hub scores for a process map following the transformation template 1 of Table 2

Activity	Hub score
Engage in negotiation	1.000
Consult [NS]	0.994
Make a visit	0.976
Host a visit	0.967
Praise or endorse	0.956

Table 4 Activities with high authority scores for a process map following the transformation template 6 of Table 2, and the thematic "CONSULT"

Activity	Authority score
USA-GRC	1.000
GRC-BEL	0.947
DEU-GRC	0.864
EUR-BEL	0.827
ITA-GRC	0.821

there are two issues pertaining to the process-like representation. The first issue is about the expected variability of flows, and the second is about the lack of definitive start and end points in intermediate relationships.

The former issue is rather common in many processes of "flexible" environments, like healthcare or customer service. A typical way to deal with it is *trace clustering*, namely to group cases, and discover a distinct process model per group, thus delivering more comprehensible results. Nevertheless, several trace clustering techniques have been proposed, varying in the representation of traces, the distance/similarity measure employed, and the cluster approach (Thaler et al. 2015), making the selection of the appropriate technique a challenging task for analysts. The critique of the existing trace clustering techniques is out of the scope of this paper, yet we present a visual approach that can support analysts in comparing their output.

Assuming that following a trace clustering technique we are provided with a hierarchical clustering tree (and thus a dendrogram that captures the tree structure), we can calculate the cophenetic distance between two traces as the height of the dendrogram at which those two traces are first joined (Sokal and Rohlf 1962). In Fig. 3 we have computed the Spearman correlation between any two cophenetic distance matrices of the trees that resulted from clustering traces using the following distance measures (Studer and Ritschard 2015):

- om: Optimal matching edit distance with a substitution-cost matrix derived from the transition rates between events
- omloc: Localized optimal matching edit distance
- nms: Distance based on number of matching subsequences
- dref1: Normalized longest common subsequence distance to the most frequent sequence
- Chi^2: Chi-squared distance over the maximal length of traces
- Euclid: Euclidean distance over the maximal length of traces

Figure 3 confirms that based on the distance measure, different clustering results will be derived. However, the added value of such a plot is that the analyst can easily check which distance measures yield similar results and which generate contrasting clusters. This empirical control can be further supported by illustrations. In particular, to further explore the similarity and difference between the alternative clustering options, we can turn to the visual product of a "tanglegram" (see Fig. 4). A

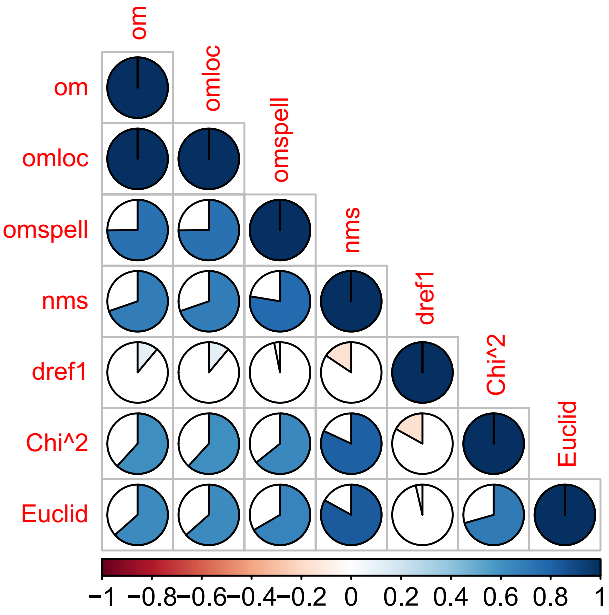

Fig. 3 Spearman correlation of cophenetic distances of the hierarchical clustering trees. The fuller the pie of a pair the greater their correlation, while the color of the pie indicate if the correlation is positive (blue) or negative (red). Illustration produced with Galili (2015) (color figure online)

tanglegram is a pair of trees on the same set of leaves with a bijection between the leaves in the two trees (Venkatachalam et al. 2010).

In a tanglegram, we expect to observe shared common sub-trees (between the dendrograms produced by the two clustering options), and distinct edges. In Fig. 4, we have used two clustering options that yield rather similar results (we based this estimation on the correlation plot of Fig. 3), and we asked for 3 clusters, which we colored in different colors. Although the produced illustration is rather dense, with a close examination we can see that there is a group of countries' pairs ("DEU-GRC", "GRC-EUR", and "EUR-GRC") that are clustered together in both options, but there are also some "important" cases like the pairs "GRC-BEL" and "EUR-BEL" that are clustered differently by the two techniques. This is an example of how analysts can increase their confidence (or their doubt) about the groupings delivered by the various techniques and visually assess the sensitivity of the clustering results. We shall note that in Fig. 4, a trace represents a sequence of events between two countries, therefore if two objects are grouped together, we can assume that the events that are happening in the two pairs of countries expose a regularity, an insight that would be inaccessible by an event-oriented (i.e., not a process-oriented) analysis.

The second issue, the lack of definitive start and end points, appears because of the nature of the problem, and it is harder to address. We discuss two relevant approaches that appear to have the potential to deal with it: *Trace alignment* and describing *patterns instead of models*. Trace alignment for a set of traces $\mathcal{T} = \{T_1, T_2, \dots, T_n\}$ is defined in Bose and van der Aalst (2012) as a mapping

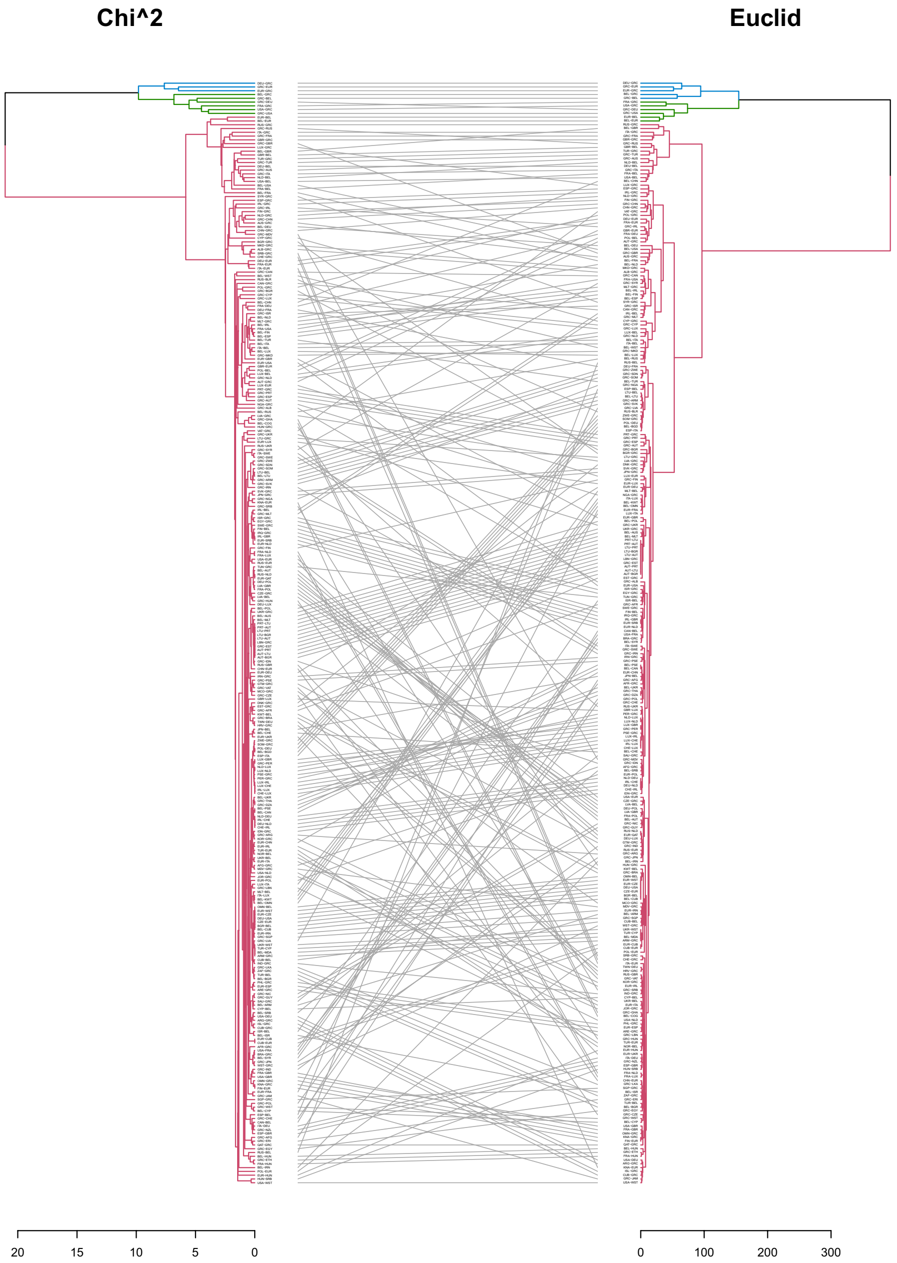

Fig. 4 A tanglegram for the hierarchical trees resulted from clustering with the Chi-squared and the Euclidean distances. Bijections connect the same objects (traces for pairs of countries), and colored branches indicate the cluster membership. Illustration produced with Galili (2015) (color figure online)

of the set of traces in $\mathcal{T}$ to another set of traces $\bar{\mathcal{T}} = \left\{\bar{T}_1, \bar{T}_2, \dots, \bar{T}_n\right\}$ where every $\bar{T}_i$ derives through T_i by the addition of one of many “gaps”, in such a way that $|\bar{T}_1| = |\bar{T}_2| = \cdots = |\bar{T}_n = m$, where m is the length of the alignment. Trace alignment is generally expected to facilitate process diagnostics, however, because of the last requirement (equal length of the aligned traces), and because in our case study the original traces varied significantly in length (from 2 to 292 events per trace), we were not able to mine any interesting results.

A different approach is to look for interesting fragments of the process, i.e., not for an end-to-end process that will demand start and end points. One option is to take a declarative approach to model business processes. Declarative techniques e.g., Maggi et al. (2011) and Pesic et al. (2007) introduce constraints in models as rules that have to be followed, but these constraints are often binary or they involve limited set of activities in pre-specified structures. To reach ampler patterns, *Local Process Models* have been proposed in Tax et al. (2016b). To generate local process models, an heuristic approach based on Markov clustering is proposed in Tax et al. (2016a).

In this work we propose a different tactic. We commend a network-based representation for portraying the flows of political event occurrences, nodes representing the events (or whatever plays the role of “activity” after the transformations of Table 2) and links between the events representing the strength of sequential occurrences in traces.

A common practice to construct such networks is to aggregate the pairwise connections in traces, and put them as edges’ weights. This practice assumes the Markov property (first-order dependency). This assumption implies that the occurrence of one event depends only on the previous event that has occurred, so it misses to capture important information, like for instance what has happened two or three steps before, because, by definition, the flow on the network is aggregated in every step (Xu et al. 2016). Therefore, we claim that higher order networks yield more appropriate representations. In Fig. 5, we plot a network of order 3 as the traces’ set fit to a Markov chain. To estimate the chain of order 3, we solved the linear programming problem of Eq. 1.

$$
\begin{aligned}
\text{minimize} \quad & \left\| \sum_{i=1}^{3} \mathcal{E} - \lambda_i M_i \mathcal{E} \right\| \\
\text{subject to} \quad & \sum_{i=1}^{3} \lambda_i = 1 \\
& \lambda_i \geq 0
\end{aligned}
\tag{1}
$$

where $\mathcal{E}$ is the distribution of events, M_i is a non-negative $m \times m$ transition matrix (with m being the number of nodes) and columns sums equal to one, and λ_i is the weight for each lag i. This optimization model is described in Ching et al. (2013) and in this work we used the implementation of Scholz (2016). We should stress that what we promote here is the visual aid of higher-order Markov chains and not their suitability for process discovery, which is left as future work.

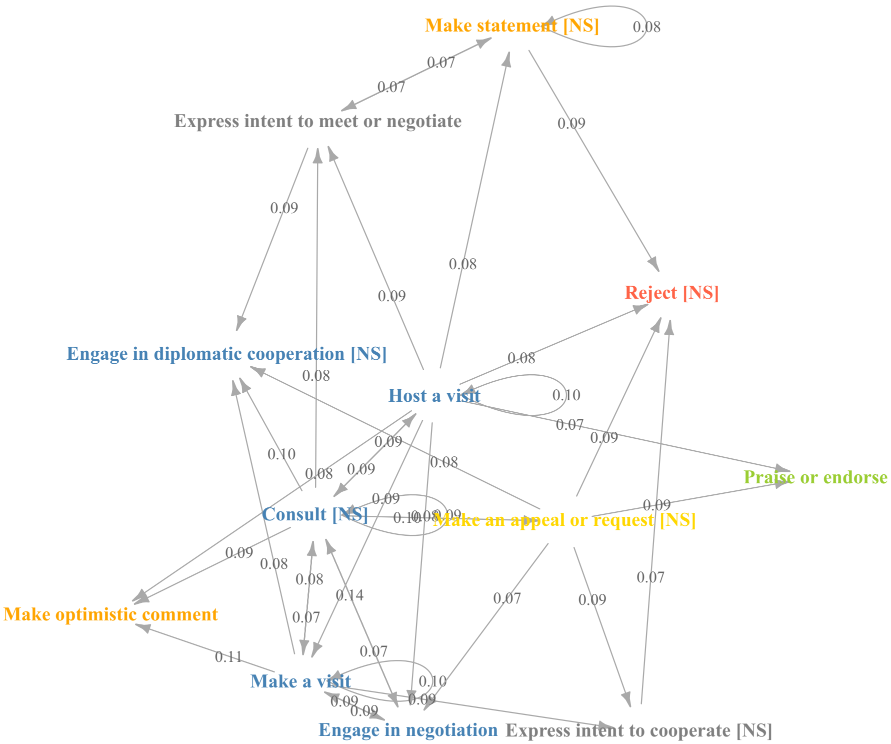

Fig. 5 Markov chain of order 3. The color of the event nodes is after their thematic characterization according to CAMEO. Only the top-12 most frequent events are showing. The visualization was created with Scholz (2016) (color figure online)

We shall notice two interesting remarks from the fit of the higher-order Markov chain to the traces' set:

- The "Reject [NS]" event appears to be a rather digressive, rambling event in Fig. 1, since it is connected with weak links to just two events, downstream to "Express intent to meet or negotiate" and upstream to "Make a visit". However, in Fig. 5, we observe that the occurrence of a "Reject [NS]" is a rather probable event, if we allow three steps (i.e., three interactions among a pair of countries) to happen.
- "Consult [NS]" is a very probable event, signifying that eventually, two countries will be involved in a "consult" event, if we allow some steps to pass.

These two insights reflect the actual situation of that time (empirically witnessed), when although we observed countries to exhibit "political correctness" and discuss issues in an institutionalized context, eventually, some actions of some country (commonly Greece), were rejected.

5 Discussion

In this work, we tried to exploit the richness and availability of raw event data by adding analytics capabilities. These capabilities are allowed by the activation of a process perspective for the (otherwise disjointed) events. Through a refinement of a process mining methodology and following the decision support paradigm, we were able to provide analysts with a guide on how to get richer insights, and with several recommendation on how to exploit the relevant visualizations to facilitate the interpretation of the events' occurrences.

There are of course several limitations in our approach. First, one must endorse our epistemological and methodological assumptions. Second, there is a plethora of process mining techniques that were not checked for their relevance to our approach. Indeed, our recommendations are relevant only in the context of the described case study, and although there are some potentials for generalization, we did not evaluate their reliability and validity for more general situations. We can not stress enough that techniques that were originally developed for a business context (the archetype motivation of Business Process Management) are not straightforwardly applicable to social/political event analysis. Anyway, this paper puts forward a big promise for event analytics, and many challenges may appear, nevertheless, given the efforts that have already been devoted to data collection issues, the focus needs to be shifted towards the analysis side.

Acknowledgements We would like to thank our graduate students Zafeiris Papavaritis and Christianna Pantermali who spent many hours in checking every event of the original dataset for relevance, and who manually filtered them out.

References

Aalen O, Borgan O, Gjessing H (2008) Survival and event history analysis: a process point of view. Springer, Berlin

Adriansyah A, Buijs JCAM (2012) Mining process performance from event logs: the BPI challenge 2012. Case Study BPM Center Report BPM-12-15. BPMcenter.org

Best RH, Carpino C, Crescenzi MJ (2013) An analysis of the TABARI coding system. Confl Manag Peace Sci 30(4):335–348

Bose RJC, van der Aalst WM (2009) Context aware trace clustering: towards improving process mining results. In: SDM, SIAM, pp 401–412

Bose RJC, van der Aalst WM (2012) Process diagnostics using trace alignment: opportunities, issues, and challenges. Information Systems 37(2):117–141 (Management and engineering of process-aware information systems)

Broström G (2012) Event history analysis with R. CRC Press, Boca Raton

Celonis (2017) Academic cloud. https://academiccloud.celonis.com. Accessed 25 Sept 2017

Ching WK, Huang X, Ng MK, Siu TK (2013) Higher-order markov chains. Springer, Boston, pp 141–176

De Leoni M, van der Aalst WM, Dees M (2014) A general framework for correlating business process characteristics. In: International conference on business process management, Springer, pp 250–266

Delias P, Kazanidis I (2017) Process analytics through event databases: potentials for visualizations and process mining. In: Linden I, Liu S, Colot C (eds) Decision support systems VII. Data, information and knowledge visualization in decision support systems, vol 282, Springer International Publishing, Cham, pp 88–100. https://doi.org/10.1007/978-3-319-57487-5_7

Delias P, Doumpos M, Matsatsinis N (2015a) Business process analytics: a dedicated methodology through a case study. EURO J Decis Process 3(3–4):357–374. https://doi.org/10.1007/s40070-015-0050-4

Delias P, Grigori D, Mouhoub ML, Tsoukias A (2015b) Discovering characteristics that affect process control flow. In: Decision support systems IV—information and knowledge management in decision processes, Springer, pp 51–63

Fails JA, Karlson A, Shahamat L, Shneiderman B (2006) A visual interface for multivariate temporal data: finding patterns of events across multiple histories. In: 2006 IEEE symposium on visual analytics science and technology, IEEE, pp 167–174

Galili T (2015) dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. Bioinformatics 31:3718–3720

Gerner DJ, Schrodt PA, Francisco RA, Weddle JL (1994) Machine coding of event data using regional and international sources. Int Stud Q 38(1):91–119

Gerner DJ, Schrodt PA, Yilmaz O, Abu-Jabr R (2002) Conflict and mediation event observations (cameo): a new event data framework for the analysis of foreign policy interactions. International Studies Association, New Orleans

Glaser BG (1978) Theoretical sensitivity: advances in the methodology of grounded theory. Sociology Press, Mill Valley (oCLC: 926199357)

Gotz D, Stavropoulos H (2014) DecisionFlow: visual analytics for high-dimensional temporal event sequence data. IEEE Trans Vis Comput Graph 20(12):1783–1792

Gotz D, Wongsuphasawat K (2012) Interactive intervention analysis. In: AMIA annual symposium proceedings, American Medical Informatics Association, Washington, DC, USA 2012, pp 274–280

Gotz D, Wang F, Perer A (2014) A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. J Biomed Inf 48:148–159

Günther CW, Rozinat A, van der Aalst WM (2009) Activity mining by global trace segmentation. In: International conference on business process management, Springer, pp 128–139

Gupta A, Jain R (2011) Managing event information: modeling, retrieval, and applications. Synth Lect Data Manag 3(4):1–141

Jiang L, Mai F (2014) Discovering bilateral and multilateral causal events in GDELT. In: International conference on social computing, behavioral-cultural modeling, and prediction

Keertipati S, Savarimuthu BTR, Purvis M, Purvis M (2014) Multi-level analysis of peace and conflict data in GDELT. In: Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis, ACM, p 33

Kwak H, An J (2016) Two tales of the world: Comparison of widely used world news datasets GDELT and EventRegistry. arXiv preprint arXiv:1603.01979

Leetaru K, Schrodt PA (2013) GDELT: global data on events, location and tone, 1979–2012. resreport, International Studies Association, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, Champaign, USA. http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf. Accessed 25 Sept 2017

Liu Z, Wang Y, Dontcheva M, Hoffman M, Walker S, Wilson A (2017) Patterns and sequences: interactive exploration of clickstreams to understand common visitor paths. IEEE Trans Vis Comput Graph 23(01):321–330

Maggi FM, Mooij AJ, van der Aalst WM (2011) User-guided discovery of declarative process models. In: 2011 IEEE symposium on computational intelligence and data mining (CIDM), IEEE, pp 192–199

Mannhardt F, de Leoni M, Reijers HA, van der Aalst WM, Toussaint PJ (2016) From low-level events to activities-a pattern-based approach. In: International conference on business process management, Springer, pp 125–141

Martjushev J, Bose RJC, van der Aalst WM (2015) Change point detection and dealing with gradual and multi-order dynamics in process mining. In: International conference on business informatics research, Springer, pp 161–178

McClelland CA (1961) The acute international crisis. World Polit 14(01):182–204

McClelland CA (1976) World event/interaction survey codebook. ICPSR, Ann Arbor

Nguyen H, Dumas M, La Rosa M, Maggi FM, Suriadi S (2014) Mining business process deviance: a quest for accuracy. In: OTM confederated international conferences "On the move to meaningful internet systems", Springer, pp 436–445

Nguyen H, Dumas M, ter Hofstede AH, La Rosa M, Maggi FM (2016) Business process performance mining with staged process flows. In: International conference on advanced information systems engineering, Springer, pp 167–185

O'Brien SP (2010) Crisis early warning and decision support: contemporary approaches and thoughts on future research. Int Stud Rev 12(1):87–104

Pesic M, Schonenberg H, van der Aalst WM (2007) Declare: full support for loosely-structured processes. In: Enterprise distributed object computing conference, 2007. EDOC 2007. 11th IEEE international, IEEE, pp 287–287

Peuquet DJ, Robinson AC, Stehle S, Hardisty FA, Luo W (2015) A method for discovery and analysis of temporal patterns in complex event data. Int J Geogr Inf Sci 29(9):1588–1611

Phua C, Feng Y, Ji J, Soh T (2014) Visual and predictive analytics on singapore news: experiments on GDELT, wikipedia, and ^sti. CoRR arXiv:1404.1996

Roy B (1994) On operational research and decision aid. Eur J Oper Res 73(1):23–26

Scholz M (2016) R package clickstream: analyzing clickstream data with markov chains. J Stat Softw 74(4):1–17

Sokal RR, Rohlf FJ (1962) The comparison of dendrograms by objective methods. Taxon 11(2):33

Song M, Günther CW, van der Aalst WM (2008) Trace clustering in process mining. In: International conference on business process management, Springer, pp 109–120

Studer M, Ritschard G (2015) What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. J R Stat Soc Ser A 179(2):481–511

Tax N, Sidorova N, van der Aalst WM, Haakma R (2016a) Heuristic approaches for generating local process models through log projections. In: 2016 IEEE symposium series on computational intelligence (SSCI), IEEE

Tax N, Sidorova N, Haakma R, van der Aalst WM (2016b) Mining local process models. J Innov Digit Ecosyst 3(2):183–196

Thaler T, Ternis SF, Fettke P, Loos P (2015) A comparative analysis of process instance cluster techniques. In: Wirtschaftsinformatik proceedings 2015, Osnabrück, pp 423–437

van Beest NR, Dumas M, García-Bañuelos L, La Rosa M (2015) Log delta analysis: interpretable differencing of business process event logs. In: International Conference on Business Process Management, Springer, pp 386–405

van Dongen B, Weber B, Ferreira D, De Weerdt J (2013) Proceedings of the 3rd business process intelligence challenge (co-located with 9th international business process intelligence workshop, BPI 2013, Beijing, China, August 26, 2013)

van der Aalst WM (2016) Process mining: data science in action, 2nd edn. Springer, Berlin. https://doi.org/10.1007/978-3-662-49851-4

van der Aalst WM, Schonenberg MH, Song M (2011) Time prediction based on process mining. Inf Syst 36(2):450–475

van der Aalst WM, Adriansyah A, van Dongen B (2012) Replaying history on process models for conformance checking and performance analysis. Wiley Interdiscip Rev Data Min Knowl Discov 2(2):182–192

van der Aalst WM, Low WZ, Wynn MT, ter Hofstede AH (2015) Change your history: learning from event logs to improve processes. In: 2015 IEEE 19th international conference on computer supported cooperative work in design (CSCWD), IEEE, pp 7–12

van der Heijden T (2012) Process mining project methodology: developing a general approach to apply process mining in practice. Master Thesis, Technische Universiteit Eindhoven, Eindhoven. http://alexandria.tue.nl/extra2/afstversl/tm/van_der_Heijden_2012.pdf. Accessed 25 Sept 2017

Venkatachalam B, Apple J, St John K, Gusfield D (2010) Untangling tanglegrams: comparing trees by their drawings. IEEE/ACM Trans Comput Biol Bioinform 7(4):588–597

Vrotsou K, Johansson J, Cooper M (2009) Activitree: interactive visual exploration of sequences in event-based data using graph similarity. IEEE Trans Vis Comput Graph 15(6):945–952

Ward MD, Beger A, Cutler J, Dickenson M, Dorff C, Radford B (2013) Comparing GDELT and ICEWS event data. Analysis 21:267–297

Wiesche M, Jurisch MC, Yetton PW, Krcmar H (2017) Grounded theory methodology in information systems research. MIS Q 41(3):685–701

Wongsuphasawat K, Gotz D (2012) Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. IEEE Trans Vis Comput Graph 18(12):2659–2668

Wongsuphasawat K, Plaisant C, Taieb-Maimon M, Shneiderman B (2012) Querying event sequences by exact match or similarity search: design and empirical evaluation. Interact Comput 24(2):55–68

Xu J, Wickramarathne TL, Chawla NV (2016) Representing higher-order dependencies in networks. Sci Adv 2(5):e1600028

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.