Scientometrics (2020) 123:677–705
https://doi.org/10.1007/s11192-020-03391-y

Scholarly event characteristics in four fields of science: a metrics-based analysis

Said Fathalla[1,2] · Sahar Vahdati[3] · Christoph Lange[4,5] · Sören Auer[6,7]

Received: 21 March 2019 / Published online: 28 February 2020
© The Author(s) 2020

Abstract

One of the key channels of scholarly knowledge exchange are scholarly events such as conferences, workshops, symposiums, etc.; such events are especially important and popular in Computer Science, Engineering, and Natural Sciences. However, scholars encounter problems in finding relevant information about upcoming events and statistics on their historic evolution. In order to obtain a better understanding of scholarly event characteristics in four fields of science, we analyzed the metadata of scholarly events of four major fields of science, namely Computer Science, Physics, Engineering, and Mathematics using Scholarly Events Quality Assessment suite, a suite of ten metrics. In particular, we analyzed renowned scholarly events belonging to five sub-fields within Computer Science, namely World Wide Web, Computer Vision, Software Engineering, Data Management, as well as Security and Privacy. This analysis is based on a systematic approach using descriptive statistics as well as exploratory data analysis. The findings are on the one hand interesting to observe the general evolution and success factors of scholarly events; on the other hand, they allow (prospective) event organizers, publishers, and committee members to assess the progress of their event over time and compare it to other events in the same field; and finally, they help researchers to make more informed decisions when selecting suitable venues for presenting their work. Based on these findings, a set of recommendations has been concluded to different stakeholders, involving event organizers, potential authors, proceedings publishers, and sponsors. Our comprehensive dataset of scholarly events of the aforementioned fields is openly available in a semantic format and maintained collaboratively at OpenResearch.org.

Keywords Scholarly events · Metadata analysis · Scholarly communication · Publishing paradigms · Metrics suite · OpenResearch.org

✉ Said Fathalla
fathalla@cs.uni-bonn.de

Extended author information available on the last page of the article

Introduction

Publishing research results in textual form is the traditional knowledge exchange paradigm of scholarly communication. Researchers of different communities use different channels for publishing. The integration and harmonization between these channels are based on the grown culture of any particular community and community-defined criteria for analyzing the quality of these channels. For example, some fields, such as medical science, use publishing in journals as the main and most valuable channel, however, some other fields, such as Computer Science,[1] publish mostly in events. Furthermore, community-defined criteria distinguish highly ranked instances of any particular class of channels as well as popular events and journals. Nevertheless, such criteria are not standardized nor centralized but generally, are transferred in a subjective way from seniors to juniors. However, a systematic and objective analysis of metadata supports researchers in better dissemination of results to the right communities. Aggregation of metadata from several data repositories, digital libraries, and scholarly metadata management services enables comprehensive analysis and services to the users of such services. Therefore, we propose the Scholarly Events Quality Assessment (SEQA) suite, a suite of metrics for common criteria, which can be used to study the various characteristics of scholarly events in different fields of science.

Besides general criteria such as the acceptance rate of events or bibliographic criteria including the impact factor of a journal and the h5-index of a person, there are community-defined criteria for ranking journals and events, such as the CORE[2] ranking. Such criteria are neither standardized nor maintained by a central instance but are transferred from seniors to juniors. Analyzing metadata of scholarly events provides statistics to answer meta-research type analytics such as the following research questions:

- How important is the role of events for communities addressed in this work?
- What criteria makes an event a high-ranked target in its community?
- What are the publishing practices via scholarly events in different disciplines?

In this work, we analyze the development of scholarly knowledge dissemination in the four fields of science Computer Science (CS), Physics (PHY), Engineering (ENG) and Mathematics (MATH). In particular, we analyze renowned events in the five CS sub-fields (World Wide Web (WEB), Computer Vision (CV), Software Engineering (SE)) Data Management (DM), and Security and Privacy (SEC) targeting to answer:[3]

- What is the orientation of submissions and acceptance rates of CS events?
- How did the number of publications of a CS sub-community fluctuate?
- Are high-impact events held around the same time slot each year?
- Which countries host the most events in each CS community?

Providing statistics about the impact of scholarly objects and measurements on the quality of research have been recently accelerated (Wilsdon 2016). The availability of certain metadata restricts the objective impact measurements to metrics related to citations only. Furthermore, diversity of the meaning of impact bring challenges for the development of

[1] For clarity, we capitalize the names of all fields of science that were subject of our study.

[2] http://www.core.edu.au/.

[3] Such a sub-community analysis was not possible for other fields for lack of data.

a robust, widely accepted impact measures. This limits the scope and quality of possible evaluations. In this work, we address the following questions in terms of impact:

- What are the characteristics of scholarly events in Computer Science, Physics, Engineering, and Mathematics?
- What are the top citation impact events in Computer Science?
- How are top citation impact events assessed by ranking services?

To go beyond citation-related measurements, an extended list of metrics is required. Analyzing scholarly event metadata, such as event dates, the number of submitted and accepted articles, location, event type, and field, can help to answer such questions. In order to systematize the evaluation, we propose Scholarly Events Quality Assessment (SEQA) suite; a suite of ten generic metrics which can be jointly applied for the selected communities and have been defined (in Sect. "Scholarly events quality assessment metrics suite") and computed based on this data. This suite involves acceptance rate, continuity, geographical and time distribution, field popularity, and productivity as well as event progress ratio and rankings, based on SJR, h5-index and ranking services, such as CORE and Qualis.[4]

In our earlier work, the particular focus was on analyzing Computer Science events in terms of continuity, geographical and time distribution, field popularity, and productivity (Fathalla et al. 2017). In 2018, we extended this work by studying the characteristics of scholarly events in four research fields, i.e., Computer Science, Physics, Engineering, and Mathematics, using statistics as well as exploratory data analysis (Fathalla et al. 2018). This article extends our previous work by

- Widening the research scope by adding more research questions,
- adopting the research workflow, as we found some tasks need to be overlapped,
- Proposing a new metrics-suite (SEQA) for scholarly events' impact assessment after redefining some metrics, such as time and geographical distributions, in order to obtain better results,
- Considering four additional metrics: average acceptance rate (AAR), community popularity (CP), field productivity (FP), and progress ratio (PR),
- Expanding the analysis of the metadata of both CS and non-CS events, and
- Providing a set of recommendations for event organizers, potential authors, proceedings publishers, and sponsors,

This study aims at answering the following research questions: (RQ1) *How important are events for scholarly communication in the respective communities*?, (RQ2) *What makes an event a high-ranked target in a community*? and (RQ3) *How can scholarly events be assessed using a mix of metrics*?

This study aims to close an important gap by analyzing the importance of scholarly events in different fields of science and assessing the comparative popularity of five major CS communities, in terms of submissions and publications. The aim is to have a momentous influence on the contributions and information needs of the different stakeholders of scholarly communication: (1) *event organizers* to assess and elevate the development and impact of their events, (2) *authors* to identify renowned events to submit their research results to, (3) *proceedings publishers* to evaluate the impact of the events whose

[4] http://qualis.ic.ufmt.br/.

proceedings are being published by certain publishers, (4) *participants* to identify candidate renowned events to attend, and (5) *event sponsors* to tighten the collaboration between industry and academia.

Finally, this work provides foundations for discovery, recommendation and ranking services for scholarly events with well-defined, transparent measures. The remainder of this article is organized as follows: Sect. "Related work" gives an overview of related work. Section "Data collection and preparation" describes the data curation process and the methodology of this study. Section "Data analysis" describes the data analysis process. Section "Results and discussion" discusses the results. Finally, Sect. "Discussion and future work" concludes and outlines the future work.

Related work

The next decade is likely to witness a considerable rise in metadata analysis of scholarly events due to the mega-trend of digitization since the preparation of manuscripts, as well as the organization of scholarly events, have become considerably easier. Preliminary work on meta-analytic methods in the 1980s focused on synthesizing the results of statistical methods and analyzing statistical variables for the purpose of making sense of them (Wolf 1986; Hedges 1984). Guilera et al. (2013) presented a meta-analysis for publications in psychology in order to provide an overview of meta-analytic research activity and to show its evolution over time. Different bibliometric indicators were used, such as the number of authors per article, productivity by country, and national and international collaborations between authors. El-Din et al. (2016) presented a descriptive analysis of Egyptian publications on the Hepatitis C virus using several indicators such as the total number of citations, authors and their affiliations, publication types and the Google Scholar citation index. Bakare and Lewison (2017) investigated the Over-Citation Ratio (i.e., researchers tending to over-cite researchers from the same country) for publications from six different scientific fields based on data from Web of Science (WoS) Clarivate Analytics. This analysis was performed in 2010 on publications from 20 countries in seven different years between 1980 and 2010. The authors conclude that chemistry and ornithology have had the highest, while astronomy and diabetes have had the lowest over citation rate.

Analyzing metadata of scholarly events has received much attention in the past decade (Fathalla and Lange 2018; Agarwal et al. 2016; Fathalla et al. 2017; Barbosa et al. 2017; Jeong and Kim 2010; Fathalla et al. 2019a). In particular, many studies (Ameloot et al. 2011; Aumüller and Rahm 2011; Barbosa et al. 2017; Fathalla et al. 2017; Hiemstra et al. 2007; Nascimento et al. 2003) have assessed the evolution of a specific scientific community within Computer Science by analyzing the metadata of publications of a specific event series. While Fathalla et al. (2017) and Biryukov and Dong (2010) have analyzed different Computer Science communities according to publications of several event series, Barbosa et al. (2017) analyzed full papers published in the Brazilian Symposium on Human Factors in Computing Systems (IHC) conference series in the period 1998–2015. Hiemstra et al. (2007) analyzed the SIGIR Information Retrieval conference publications from the period 1978–2007. Biryukov and Dong (2010) investigated collaboration patterns within a research community using information about authors, publications, and conferences. Similarly, Aumüller and Rahm (2011) analyzed affiliations of database publications using author information from DBLP, and Nascimento et al. (2003) analyzed the co-authorship graph of SIGMOD conference publications. Yan and Lee (2007) proposed

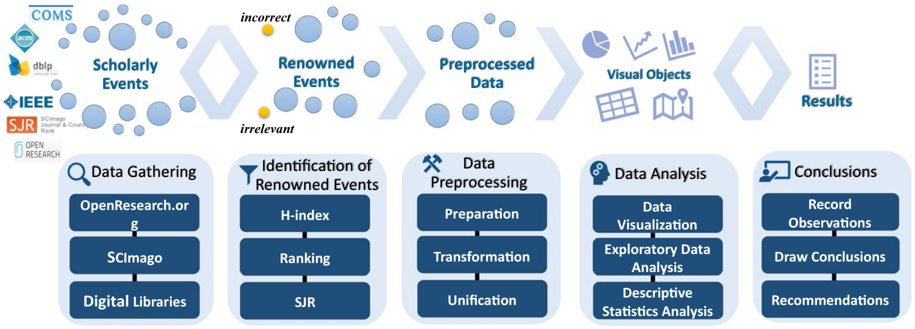


Fig. 1 The workflow of the scholarly events metadata gathering, preparation and analysis

a list of alternative measures for ranking events based on the “goodness” of the articles published in these events. In 2018, we published our scholarly events dataset, i.e., EVENTSKG Fathalla and Lange (2018), a Linked dataset representing a comprehensive semantic description of 40 renowned event series from six Computer Science communities since 1969. The EVENTSKG dataset is a new release of the previously presented dataset called EVENTS Fathalla and Lange (2018) with a 60% additional event series. Notably, it uses the Scientific Events Ontology Fathalla et al. (2019) as a reference ontology for modeling event metadata and connects related data that was not previously connected in EVENTS.

Despite these continuous efforts, a comprehensive comparative analysis of the characteristics of scholarly events in multiple scientific communities has not been performed to the best of our knowledge. None of them provides services to ease the process of gaining an overview of a field, which is the contribution of this work. We found that the characteristics of these events have not been dealt with in depth. Therefore we conducted this analysis, which is based on a comprehensive list of ten metrics, considering quality in terms of event-related metadata in six communities within CS and also events from other communities.

Data collection and preparation

This study represents a comprehensive analysis of a collected metadata about scholarly events, such as conferences, workshops, and symposiums. It has been applied to the information of four selected research communities belonging to four fields of science. Metadata analysis refers to the statistical methods used in research synthesis for drawing conclusions and providing recommendations from the obtained results. In order to conduct the analysis, a collection of data should be provided. The workflow of this study (see Fig. 1) comprises five phases:

1. Data gathering,
2. Identification of relevant and renowned events,
3. Data preprocessing,
4. Data analysis and visualization, and
5. Conclusions.

Table 1 Research fields and corresponding sub-fields

Fields	Sub-fields
Computer Science	World Wide Web (WEB), Computer Vision (CV), Software Engineering (SE), Data Management (DM), Security and Privacy (SEC), Knowledge Representation and Reasoning (KR), Computer Architecture (ARCH), Machine Learning (LRN)
Physics	Astronomy, High Energy Physics, Particle Accelerators, Applied Physics and Mathematics, Nuclear Science, Nanomaterials, Neutrino Detectors, Geophysics
Engineering	Civil Engineering, Mechanical Engineering, Chemical Engineering, Electrical Engineering
Mathematics	Algebra, Mathematical Logic, Applied Mathematics, General Mathematics, Discrete Mathematics

In the proposed workflow, several steps can be overlapped, i.e., executed in parallel. For example, in our execution, data gathering and the identification of relevant events were performed simultaneously. While collecting data, we found some renowned events that had not been considered from the beginning; therefore, we began to collect the metadata of these events as well. During these steps, a series of challenges have been encountered, such as data duplication, incomplete data, incorrect data, and the change of event titles over time. Therefore, *Data Curation* methods as a set of activities related to organization, integration, and annotation of the data (Sabharwal 2015) have been applied to ensure that the data is fit for the intended purpose, and can be reused efficiently. Data analysis and recording observations steps were also executed in parallel. Further details about each step are given in the following sections.

Data gathering

Data Gathering is the process of collecting data from a variety of online sources in an objective manner. The relevant metadata of 3,704 various scholarly events has been collected, involving conferences, workshops, symposiums, and meetings in the CS, PHY, ENG, and MATH fields. The corresponding sub-fields of each field involved in this study are listed in Table 1. Computer Science sub-fields were derived from analyzing the topics of each event series and map them to the ACM Computing Classification System (CCS),[5] while sub-fields of non-CS fields were derived from the Conference Management Software COMS.[6] We focused on the WEB, CV, SE, DM and SEC sub-fields of CS because they were the top-5 sub-fields in our datasets in terms of data availability and had the highest number of submissions in the last decade. This data, including title, series, sub-field, start date, end date, homepage, country, and h5-index, has been collected from different sources. For Computer Science, metadata of scholarly events is available as Linked Data through DBLP.[7] Multidisciplinary data harvesting services, for example, metadata of OpenAIRE (Alexiou et al. 2016) project,[8] are planned to be used in future work. Other services such

[5] https://dl.acm.org/ccs/ccs.cfm.

[6] https://www.conference-service.com/.

[7] https://dblp.uni-trier.de/.

[8] http://openaire.eu.

as the ACM Digital Library,[9] and conference city[10] are built on top of event metadata and provide search facilities.

Data harvesting sources Two major datasets are used in this study: 1) *OpenResearch dataset* (ORDS) (6000+ events) and 2) *SCImago dataset* (SCIDS) (2,200+ events). In fact, SCIDS stores metadata of each event in terms of SJR, h5-index, number of references in each paper and the number of citations for each event's proceedings volume. On the other hand, ORDS stores different attributes, such as start date, end date, number of submissions and number of publications. Therefore, we can use different statistical methods and consequently obtain different results. The reason for collecting data from two separate sources is that CS communities, compared to other research fields, archive more information about past events, such as acceptance rate, location, date, and the number of submitted and accepted papers. Furthermore, there are many online services for archiving past events metadata and ranking their proceedings, such as DBLP, and AMiner.[11]

OpenResearch.org (OR) platform[12] is an additional source of semantically structured data, which also supports data analysis. At the time of writing, OR contains crowd-sourced metadata about more than 66,000 conferences, +1000 workshops, and +370 event series. OR is based on SemanticMediaWiki (SMW) and supports researchers in collecting, organizing and disseminating information about scholarly events, tools, projects, people and organizations in a structured way (Vahdati et al. 2016). Information is represented in wiki pages with semantic annotations, with the possibility to be queried via a SPARQL endpoint as well as inline queries embedded into wiki pages. For an exchange with other tools, metadata about a given event series can be exported from OR into the CSV format using the following SMW ASK query.

```
{{#ask: [[Event in series::{{PAGENAME}}]]
| ?Submitted papers                    | ?Accepted papers
| ?Field                               | ?Acceptance rate
| ?Attendees                           | ?Start date
| ?Has location country                | ?Has program chair
| sort = ,                             | order = asc
| format = csv
}}
```

SCImago is an online database that contains information about journals and event proceedings since 1996 for 27 different research fields, including Computer Science, Physics, Engineering, and Mathematics. Journals or event proceedings can be grouped by research field, sub-fields or by country (Scimago 2007).

Identification of renowned events

To identify renowned events to be used in this study, we used the following metrics, which are commonly used to identify high-quality events in each field of science.

[9] http://dl.acm.org/.

[10] http://www.conference.city/.

[11] https://aminer.org/ranks/conf.

[12] http://openresearch.org.

Google Scholar Metrics (GSM)[13] provides ranked lists of conferences and journals by scientific field based on a 5-year impact analysis over the Google Scholar citation data. The top-twenty ranked events and journals are shown for each (sub-)field. The ranking is based on the two metrics h5-index[14] and h5-median.[15] *Qualis (Q)* uses h-index as a performance measure for conferences. Based on the h-index percentiles, the conferences are grouped into performance classes that range from A1 (best), A2, B1, ..., B5 (worst). The main limitation of GSM, ranking only the top 20 venues (including both journals and events) in each field, prompted us to find alternatives, including historical and comprehensive impact factors.

SCImago Journal Rank (SJR indicator) is a measure of the scientific influence of scholarly journals and events based on both the number of citations received by a journal and the prestige of the journals where such citations come from González-Pereira et al. (2010). It is publicly available via an online portal.[16] This rank is calculated based on the information contained in the Scopus[17] database starting from the year 1996.

CORE (*Computing Research and Education Association of Australasia*) provides a ranking method for major conferences in computing. The ranking is determined by a mix of indicators, including citation rates, paper submission, acceptance rates and the visibility and research track record of the key people hosting and managing the conference. Based on these metrics an event can be categorized into six classes A*, A, B, C, Australian, and unranked. The portal shows international event series in the first four classes.

Data preprocessing

The main objective of the data preprocessing phase is to (1) fill in missing data, (2) identify incorrect data, (3) eliminate irrelevant data, and (4) resolve inconsistencies. In order to prepare the raw data for analysis, we carried out four preprocessing tasks: *data integration*, *data cleansing*, *data transformation* and *name unification*.

(1) *Data integration* involves combining data from multiple sources into meaningful and valuable information. In addition, this process also involves eliminating redundant data, which might occur during the integration process.
(2) *Data cleansing* focuses on curation of incorrect or inaccurate records. For instance, some websites provide incorrect information about events' submissions and accepted papers. We verified this information against the official websites of the events or proceedings published in digital libraries.
(3) *Data structure transformation* involves converting cleaned data values from unstructured formats into a structured one. For instance, data collected from websites of the events as text (i.e., unstructured format) is manually converted to CSV (i.e., structured format) and subsequently to RDF.
(4) *Name unification* involves integrating all individual events of a series with multiple names under its most recent name. This is important for the researchers who want to submit their work to know the recent name rather than the name that had been in use

[13] https://scholar.google.com/intl/en/scholar/metrics.html.
[14] h5-index is the h-index for articles published in the last 5 complete years.
[15] h5-median is the median number of citations for those articles in the h5-index.
[16] http://www.scimagojr.com.
[17] https://www.scopus.com/.

Table 2 Event name unification: overview of name changes for popular events

Unified acronym	Acronym	Time span	Full title
TheWeb	WWW	1994–2017	World Wide Web Conference
	TheWeb	2018–2019	The Web Conference
VR	VRAIS	1993–1998	Virtual Reality Annual International Symposium
	IEEE VR	1999–2018	IEEE Virtual Reality
ASE	KBSE	1990–1996	Knowledge-Based Software Engineering Conference
	ASE	1997–2018	Automated Software Engineering
ISWC	SWWS	2001	Semantic Web Working Symposium
	ISWC	2002–2018	International Semantic Web Conference
FOCS	SWCT	1960–1965	Symposium on Switching Circuit Theory and Logical Design
	SWAT	1966–1974	Annual Symposium on Switching and Automata Theory
	FOCS	1975–2018	Annual Symposium on Foundations of Computer Science
ISMAR	IWAR	1999	International Workshop on Augmented Reality
	ISAR	2000–2001	International Symposium on Augmented Reality
	ISMAR	2002–2018	International Symposium on Mixed and Augmented Reality
ISSAC	EUROSAM	1974, 1979, 1982, 1984	International Symposium on Symbolic and Algebraic Manipulation
	SYMSAC	1971, 1976, 1981, 1986	Symposium on Symbolic and Algebraic Manipulation
	ISSAC	1988–2018	International Symposium on Symbolic and Algebraic Computation
SPLASH	OOPSLA	1986–2009	Conference on Object-Oriented Programming, Systems, Languages, and Applications
	SPLASH	2010–2018	Systems, Programming, Languages and Applications: Software for Humanity
TAM	TAM	2011–2014	Transport Asset Management
	ICE-TAM	2015–2018	Institution of Civil Engineers:Transport Asset Management
AWF	AWF	1987–2009	Annual Workshop on Automorphic Forms and Related Topics
	AWF	2010–2018	Automorphic Forms Workshop
P-ADIC	P-ADIC	1990–2014	International Conference on p-Adic Functional Analysis
MEMS	MEMS	1989–1990	Conference on An Investigation of Micro Structures, Sensors, Actuators, Machines and Robots
	MEMS	1991–2018	International Conference on Micro Electro Mechanical Systems

for the longest time, as shown in Table 2. The rationale for name unification is that we observed that some events have changed their names once or several times since they had been established. The change sometimes happens because of changing the scale of the event to a larger scale, e.g., from Symposium to Conference or from Workshop to Symposium, for example, ISWC and ISMAR respectively. Also, the change sometimes happens because of adding a new scope or topic, such as SPLASH. In addition, conferences such as SPLASH keep both names, the old and the new one. In this case, we also keep the most recent name. This led us to perform a unification process before beginning to analyze the data. For example, IEEE VR, or just VR is the unified name of IEEE Virtual Reality, which assumed this name in 1999 and was previously known as VRAIS from 1993 to 1998. Notably, we observed that ISSAC was alternatively changing its name since 1966. The name "ISSAC" was first used, and still used till 2019, for the 1988 meeting, when the European EUROSAM/ EUROCAL/ EUROCAM conferences were merged with the North American SYMSAM/ SYMSAC stream.

Data analysis

The heart of this work is the exploratory analysis of the metadata for the selected events over the past 30 years. Generally, the data analysis process is divided into three categories: Descriptive Statistics Analysis (DSA), Exploratory Data Analysis (EDA) (Martinez et al. 2010), and Confirmatory Data Analysis (CDA). DSA describes the main aspects of the data being analyzed. EDA is an approach for data analysis that explores new features and unknown relationships in the data to provide future recommendations and CDA aims to understand data from a hypothesis-testing perspective to confirm or falsify existing hypotheses (Behrens and Yu 2003). The analysis presented in this work is based only on DSA and EDA because our purpose is to describe and explore new insights.

Analysis tools

The methodology and tools used in this work are based on our preliminary study on this topic, which only considered the Computer Science community (Fathalla et al. 2018). We first defined analysis metrics, then chose suitable tools for computing them and evaluating the results of the computation. We defined statistical metrics over numeric values, as well as metrics having other complex datatypes, focusing on conferences because of their high impact on research communities. We chose spreadsheets as the main tool to compute statistical metrics over numeric values; the evaluation of the results is supported by charts. OpenResearch.org provides further components for visual analytics (such as Calendar View, Timeline View, and Map View), in particular for displaying non-numeric results (e.g., the conferences with the highest number of submissions). Even though spreadsheets are, in principle, based on the relational data model, they practically lack support for joins across sheets. Joins may be required for connecting information about events to information about related entities, such as persons participating in events. Therefore, a set of queries are designed to support this challenge. The SPARQL query language for RDF, which is supported by *OpenResearch.org* through its public SPARQL endpoint, facilitates such join computations. However, while SPARQL also supports fundamental statistical analysis via aggregate functions, this type of analysis is better supported by spreadsheets. For example, finding all events related to Computer Security, which took place in Europe[18] along with their acceptance rate (less than 20%) in the last five years; this requires joins between field/topic and event entities. The following query is designed for this purpose:

```
SELECT ?event ?endDate ?startDate ?city ?country ?wikipage ?acceptanceRate ?continent
WHERE {
  ?e              rdfs:label                       ?event .
  ?e              a                                category:Computer_Security .
  ?e              icaltzd:dtstart                  ?startDate .
  ?e              icaltzd:dtend                    ?endDate .
  ?e              property:Acceptance_rate         ?acceptanceRate .
  ?e              swivt:page                       ?wikipage .
  ?e              property:Has_location_country    ?country .
  ?country        rdfs:subClassOf                  ?partContinent .
  ?partContinent  rdfs:subClassOf                  ?continent .
  ?continent      rdfs:isDefinedBy                 site:Category:Europe .
  FILTER (
    ?acceptanceRate <  20.0 &&
    ?startDate      >= "2013-01-01"^^xsd:date &&
    ?endDate        <  "2018-01-01"^^xsd:date
  )
}
ORDER BY DESC(?acceptanceRate)
```

Scholarly events quality assessment metrics suite

We defined SEQA metrics suite, which contains ten metrics involving numeric values and complex data types, in order to study various characteristics of high-quality events in different fields of science. Compared to previous work, new metrics are used to compare scholarly events and respective their communities, such as SJR indicator, h5-index, field productivity, continuity, and progress ratio, the latter three are defined by ourselves. SEQA comprises the following metrics:

1. *Acceptance Rate (AR)* is the ratio between the number of accepted articles and the number of submitted ones for a particular event in a particular year. The average acceptance rate (AAR) is calculated for all editions of a particular series to get an overview of the overall acceptance rate of this series since the beginning.
2. *Continuity (C)* reflects the continuation of the conference since its beginning. We propose a formula (Eq. 1) to compute the percentage of continuity for a specific event, where C is the continuity, E is the number of editions of the event, R is the regularity of the event editions, e.g., $R = 1$ for events that take place every year, and A is the age, i.e., the number of years since the event took place first. For events that changed its regularity, e.g., from 2 to 1, we computed the continuity of both periods and the overall continuity is the average. The year is the granularity for this metric.

[18] The complexity of the relation between a country and its continent is owed to the way OpenResearch.org organizes such knowledge to provide convenient browsing by regions of continents such as “Western Europe”.

$$C = \min \left\{ 100\%, \frac{E \times R}{A} \right\} \tag{1}$$

3. *Geographical Distribution (GD)* measures the number of distinct locations, in terms of the hosting country, visited by an event. The key question is which countries hosted most of the events of a series in the dataset, and how frequently a country has hosted an event during a given period of time. Thus, we can conclude which country pays more attention to a particular type of events, in terms of the field of research. In addition, we study the change of the location of each event from year to year which denoted by ΔL_n (Eq. 2), where l_n is the location of an event in a year and $l_n + 1$ is the location of the same event in the next year.

$$\Delta L_n = \begin{cases} 1 & \text{if } l_n \neq l_{n-1} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

 Then, the mean of these changes ($\bar{x}$) is computed to measure the rate of the distribution of each event since the beginning (Eq. 3). The higher this value is for an event, the more frequently the host country of an event changed.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n-1} \Delta L_i \tag{3}$$

4. *Time Distribution (TD)* refers to the period of time each year in which the event takes place. In this study, we consider the month in which the event takes place. Then, the standard deviation (σ) is computed in order to quantify the variation or dispersion of the month in which the event takes place each year. It is important for researchers, who are interested in a particular conference, to be able to estimate when a particular event will be held next time, which helps to devise a submission plan.
5. *Community Popularity* reveals how popular an event is in a research community, in terms of the number of submissions. The more submissions to events of a particular field, the more popular is this field. This also gives an indication about which community attracts a larger number of researchers.
6. *Field Productivity (FP)* reveals how productive, in terms of the number of publications, a research field in a given year within a particular time span is. In other words, FP is the ratio of the publications of all events in a given research field in a given year to the total number of publications of all events in the same field in the dataset within a fixed-time period. The field productivity for a research field (F) in a year (y) is defined in Eq. 4, where F is modeled as the set of all events in the given field (F), $P_y(e)$ is the number of publications of all events in F which held in a given year y, and n is the number of years in the time span of the study.

$$FP_y(F) = \frac{\sum_{e \in F} P_y(e)}{\sum_{i=1}^{n} \sum_{e \in F} P_i(e)} \tag{4}$$

7. *Progress Ratio (PR)* sketches to the progress of an event in a given year within a particular time span. It is the ratio of the publications of an event in a given year to the total number of publications by all events in the same field in a given period of time. Thus, this metric gives an overview to events' organizers about the progress of their event compared to other competitive events in the same field. The progress ratio for

an event (e) in a year (y) is defined in Eq. 5, where $P_y(e)$ is the number of publications of e in y and n is the number of years in the time span of the study.

$$PR_y(e) = \frac{P_y(e)}{\sum_{i=1}^{n} P_i(e)} \tag{5}$$

8. *SJR indicator* is the average number of weighted citations received in a particular year by the papers published by an event in the three previous years (González-Pereira et al. 2010).
9. *H5-index* is the h-index for articles published in the last five complete years. It is the highest number h such that h articles published in 2013–2017 have at least h citations each.
10. *Citation count* is the number of citations papers receive, according to SCIDS.

Using these metrics, it is possible to provide a flexible and broad study on various characteristic dimensions of scholarly events in different fields of science.

Results and discussion

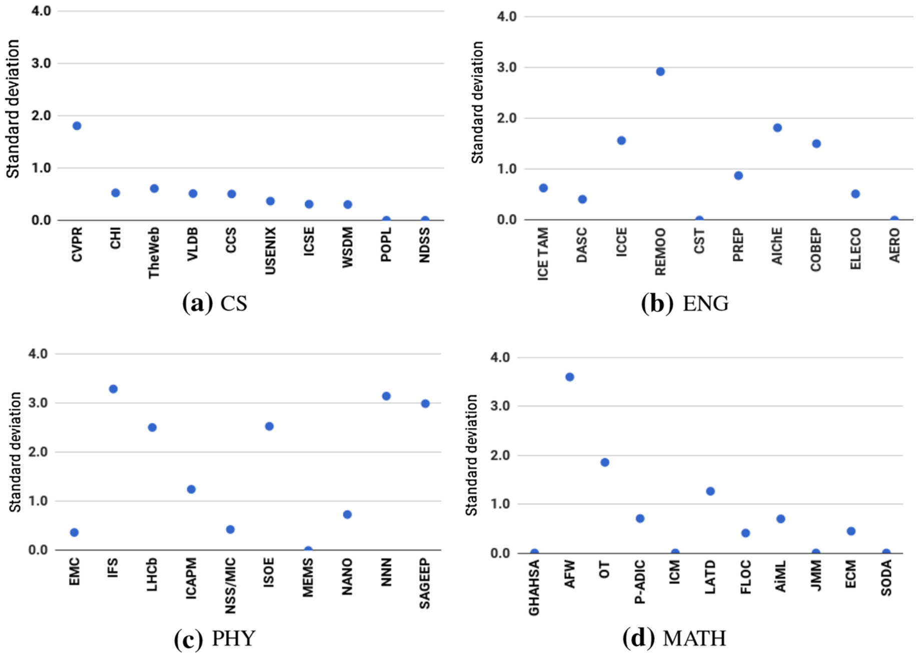
In this section, we report the results of our analysis of events metadata within the two datasets over the past 30 years according to the SEQA metrics defined in Sect. "Scholarly events quality assessment metrics suite". One notable observation is that there is no comprehensive information about the number of submissions and publications in other fields than CS. Therefore, metrics such as acceptance rate, FP and PR cannot be practically applied to events belonging to this fields. For the same reason, we categorize our results into three categories: (1) scientific fields analysis, (2) CS sub-fields analysis, and (3) individual events analysis.

Scientific fields analysis

This section presents the results of analyzing metadata of events from all considered scientific fields, i.e., CS, PHY, ENG, and MATH, with respect to the proposed metrics that can be applied, which are TD, GD, h5-index, continuity, SJR and citation count.

Time distribution We analyzed the time distribution metric, in terms of the standard deviation of the month of the year in which the event takes place, for all events of CS, MATH, ENG, and PHY in ORDS in the last two decades. Namely, EMC (International Symposium on Electromagnetic Compatibility) has been held 17 times in August and only 3 times in July, therefore, it has a low standard deviation of $\sigma = 0.4$, while NDSS (Annual Network and Distributed System Security Symposium) has been held in February every year since the beginning, therefore $\sigma = 0$. Notably, CS events have the lowest σ among events of other fields. Overall, we observed that most editions of the high-quality events in all fields have always been held around the same month every year, i.e., their time distributions have low standard deviations (Fig. 2).

Geographical distribution We analyzed the geographical distribution of all CS, MATH, PHY and ENG events in the last two decades. As shown in Fig. 3, the USA hosted 50% or more of the scholarly events in all fields during the whole period, while all other countries



(a) CS

(b) ENG

(c) PHY

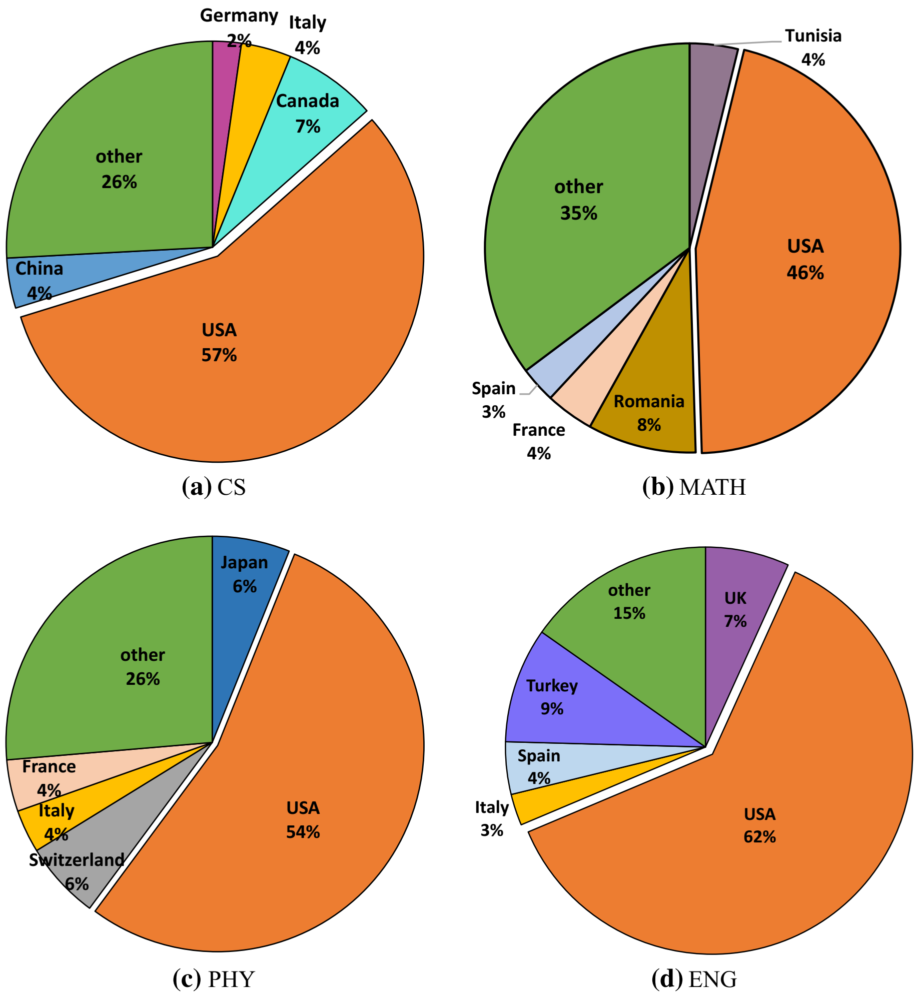
(d) MATH

Fig. 2 Time distribution of events in CS, MATH, PHY and ENG in the last two decades

have significantly lower percentages. For instance, Canada hosted only 7% of CS events and a significantly low percentage of events of the other fields, while France hosted 4% of both MATH and PHY events.

H5-index To compare the impact of events of the four scientific fields, we analyzed the h5-index of the top-25 events in each field. Figure 4 shows the frequency distribution of events by categorizing the h5-index of the events into four ranges (0–10, 11–20, 21–30 and 30+). The slices of each pie chart compare the frequency distribution of events in each field with respect to the h5-index. The CS community has the highest number of events (92%) with ($h > 30$), while the ENG community has the lowest one (16%). The number of MATH events with ($h > 30$) is as high as that of PHY, while each of them is almost twice as high as ENG. Also, the number of ENG events with ($21 \leq h < 30$) is as high as that of PHY. Overall, we found that CS has the highest number of high-impact events, while ENG has the lowest. This can be, for example, attributed to the size of the field and its sub-fields and their fragmentation degree, since a large community results in higher citation numbers. Also, it might be an indication of the importance of events for scholarly communication of this community (e.g., in comparison to journals).

Continuity As shown in Fig. 5, all events in all fields have a continuity greater than 90% except for NNN (International Workshop on Next generation Nucleon Decay and Neutrino Detectors) and ICE-TAM (Institution of Civil Engineers-Transport Asset Management), which have continuity of 88% and 86% respectively. The reason is that NNN was not held in 2003 and 2004, and ICE-TAM was not held in 2013. For CS events, the continuity of USENIX (Usenix Security Symposium) is 93% because it was held every year from 1990 except for two years (1994 and 1997). This emphasizes that even the lowest continuity value of CS events is relatively high, in comparison to the other fields. Overall,



(a) CS

(b) MATH

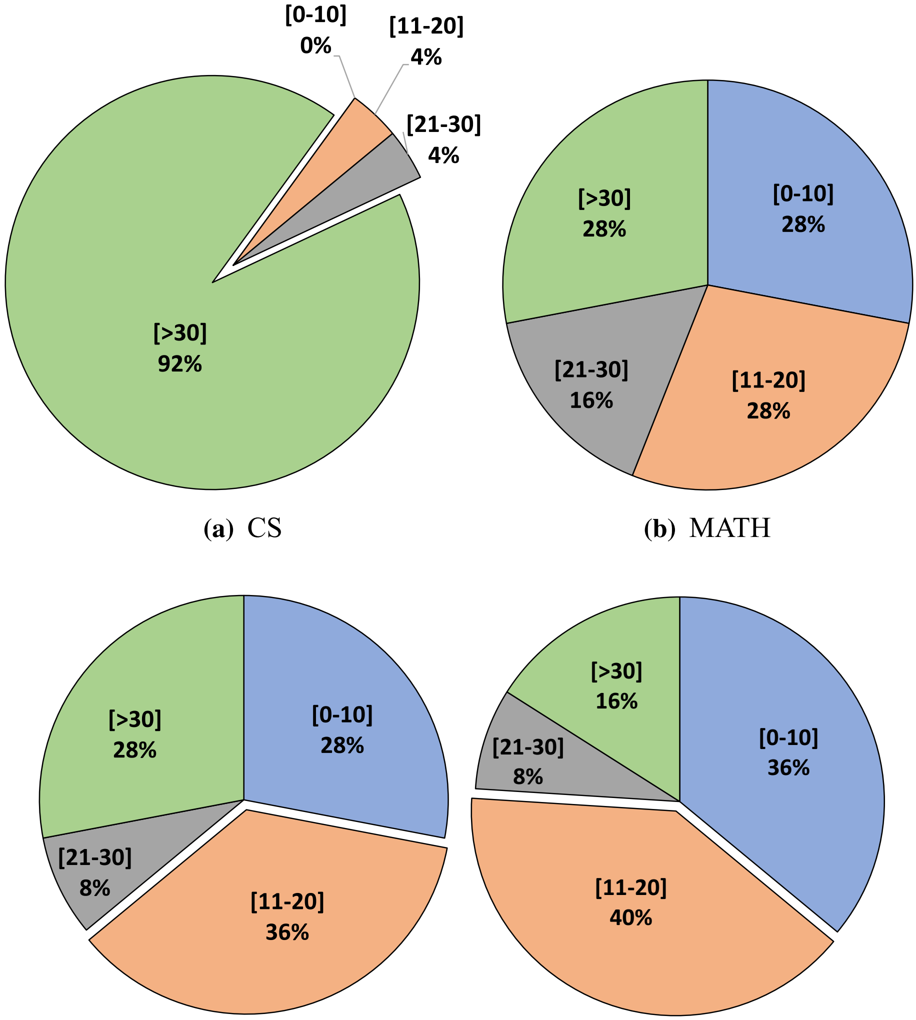
(c) PHY

(d) ENG

Fig. 3 Geographical distribution of CS, MATH, PHY and ENG events in the last two decades

we observed a very high continuity among renowned events, which is an indication of stability and of the attractiveness of hosting and organizing such events.

SJR indicator We calculated the average SJR indicator of all events, in SCIDS. As shown in Table 3, CS communities have an average SJR of 0.23 which is almost twice the value of PHY and ENG each; MATH comes next. As the SJR indicator is calculated based on the number of citations, we can infer that CS and MATH communities were more prolific or interconnected in terms of citations in 2016 compared to PHY and ENG. Since PHY had the highest number of articles published in 2016 (among other fields), but not the highest SJR indicator. This can be rather attributed to the number of citations per article is lower than in other fields. On average, a CS paper contains about 20 references (refs/paper), while a PHY paper contains only 15 references. In terms of the total number of references included in the papers published in 2016 (total refs.), CS has the highest number of references, while the ENG field has the lowest.

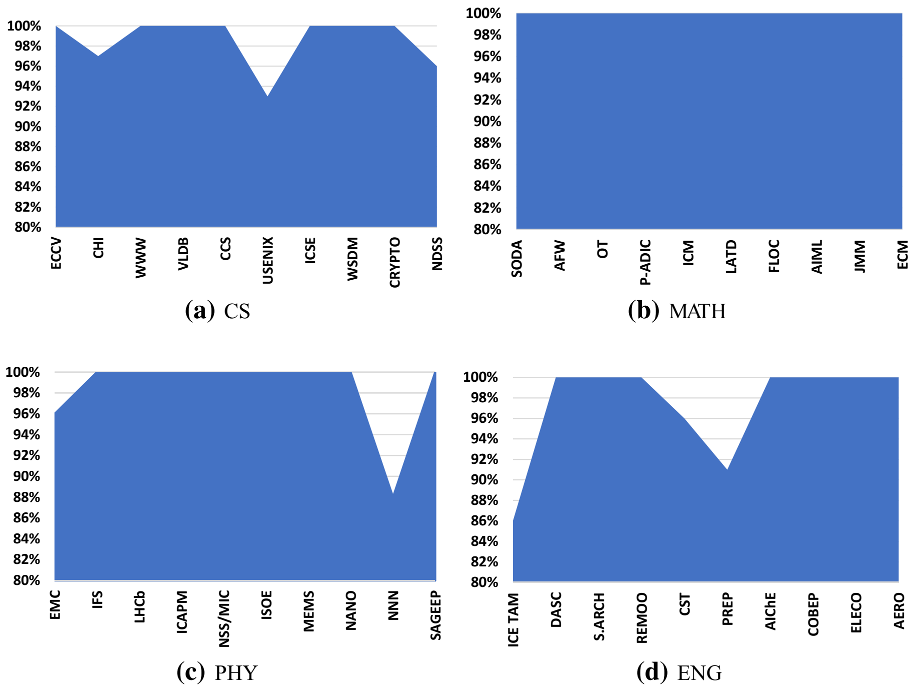


(a) CS **(b)** MATH

(c) PHY **(d)** ENG

Fig. 4 Frequency of the top-25 events in CS, MATH, PHY and ENG in terms of their h5-index

Citation count We analyzed the number of citations of all proceedings papers of events that took place in, e.g., Germany, for the CS, ENG, MATH, and PHY between 2007 and 2016. Figure 6 illustrates the development of the number of citations for each field over the period 2007–2016. This is an indication, that there is a relatively large number of researchers in Germany working in CS. While the number of citations has increased for all communities during this period, the strongest increases were observed in CS and ENG. The leading role of CS has persistently increased throughout the whole period. The citations for PHY and MATH are relatively low and are almost similar. Overall, we can see a clear upward trend in the number of citations of CS publications, compared to a slight increase in the three other fields.



(a) CS

(b) MATH

(c) PHY

(d) ENG

Fig. 5 Continuity of CS, MATH, PHY and ENG events in the last two decades

Table 3 Scientometric profile of the top CS, PHY, ENG and MATH events held in 2016

Metrics	CS	PHY	ENG	MATH
Max(h)	192	125	52	125
Avg(h)	6.58	6.65	4.09	6.79
Conf ($h > 10$)	151	28	21	25
Avg. SJR	0.23	0.14	0.14	0.21
Papers (2016)	13,234	16,795	1675	16,585
Papers (2013–2015)	163,556	90,245	46,790	68,814
Total refs. (2016)	262,548	248,216	27,137	258,275
Refs/paper	20	15	16	16

Data obtained from SCImago database

CS sub-fields analysis

This section focuses on analyzing events of five CS sub-fields (WEB, CV, SE, DM and SEC) based on the number of submissions and accepted papers, and all applicable metrics, such as AAR, FP, and PR.

Geographical distribution We analyzed the geographical distribution of the top-5 events in each CS community since 1973. As illustrated in Fig. 7, the USA hosted most editions of events in all CS communities. For instance, the USA hosted 40% (41 out of 96) of WEB events, 37.5% (54 out of 144) of CV events, 67.5% (104 out of 154) of SE events, 25.1%

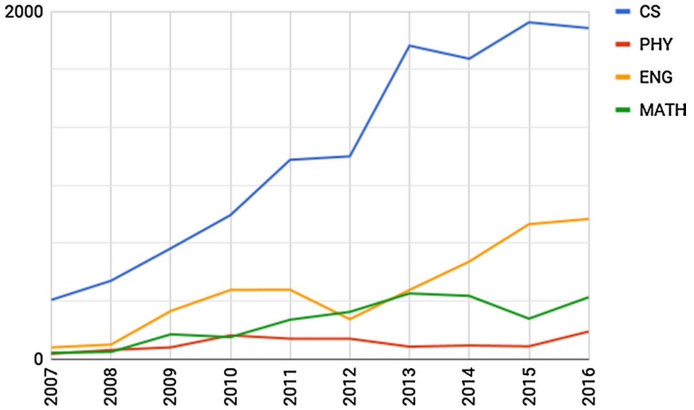
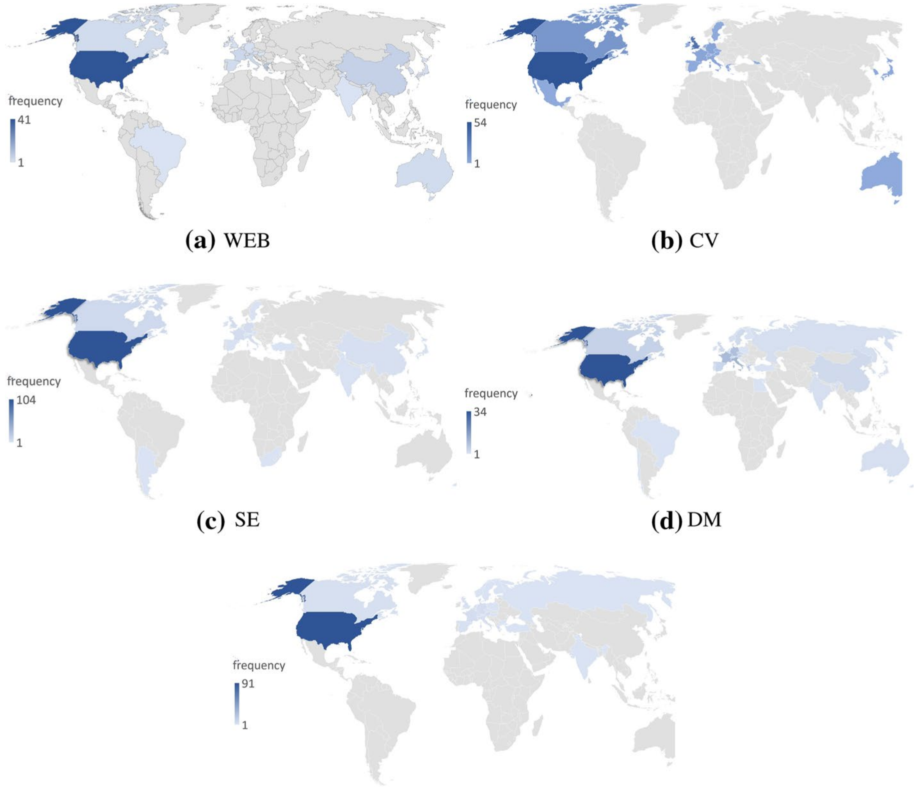


Fig. 6 Citation count by different communities in Germany



(a) WEB

(b) CV

(c) SE

(d) DM

(e) SEC

Fig. 7 Geographical distribution of the top-5 events in each CS community since 1973

(34 out of 135) of DM events, and 66.4% (91 out of 137) of SEC events. The DM community has the widest geographical distribution of events with 37 different countries hosting

Table 4 Accepted and submitted papers measures for five CS communities over three 5-years intervals

		Accepted papers			**Submitted papers**		
		2003–2007	**2008–2012**	**2013–2017**	**2003–2007**	**2008–2012**	**2013–2017**
WEB	Avg.	197	310	338	1146	1818	1905
	Min	143	264	251	921	1739	1491
	Max	223	378	507	1363	1897	2598
CV	Avg.	342	866	965	1148	3150	3914
	Min	226	593	632	1012	2312	2954
	Max	473	1177	1255	1909	4047	4901
SE	Avg.	148	211	302	958	1180	1486
	Min	116	190	261	751	1094	1405
	Max	167	237	320	1091	1290	1558
DM	Avg.	211	327	383	1279	1543	1481
	Min	176	282	195	978	1456	727
	Max	265	364	503	1727	1611	2248
SEC	Avg.	145	195	397	912	1103	1915
	Min	142	161	298	788	916	1485
	Max	152	258	508	980	1326	2353

137 events, while the WEB and SE communities have the most narrow geographical distribution with only 21 countries hosting 96 and 154 events, respectively. We observed that some events are restricted to one continent, such as EUROCRYPT, which has been held every year in Europe since 1982 and CRYPTO which has been held every year in North America since 1995. Strikingly, we observed that most of renowned events in SEC have been held in North America, particularly in the USA (83%), which indicates that the USA pays particular attention to this field. Notably, it is observed that the USA hosted most of the top-5 events in all communities.

Time distribution We observed that most editions of top event series are held around the same month each year (see Table 6). Namely, the WSDM Conference (ACM International Conference on Web Search and Data Mining) has been held every year since 2008 in February and PLDI (conference on Programming Language Design and Implementation) has been held every year since 1987 in June.

Community popularity We compared the popularity of the five CS communities in terms of the number of submissions and accepted papers (see Table 4). The CV community had the highest number of submissions and accepted papers during the three 5-year time windows. The lead of CV in terms of submissions and accepted papers has continuously increased over the whole period, i.e., 2003–2017, until reaching nearly 4000 submissions, on average, by the end of 2017 (highlighted in yellow). For example, the number of submitted papers in the period 2008–2012 (3,150 papers) is twice as large as of the period of 2003–2007 (1,148 papers) (highlighted in gray). Submissions, as well as accepted papers of the WEB community, have gradually increased throughout the whole time span. The submissions of SEC have doubled in the last five years and consequently, the accepted papers (highlighted in green). Differently, we observed that the average number of submitted papers of the DM community has slightly decreased in the last period, while the average number of submitted papers has slightly increased (highlighted in red). Overall, the CV community has had most submissions among the CS communities, while DM had the least.

Field productivity The slices of the pie chart in Fig. 8 compare the cumulative field productivity of eight CS communities in the last ten years. We only applied the FP

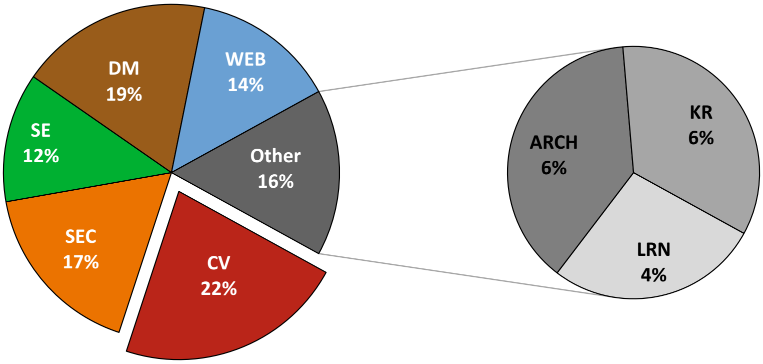


Fig. 8 Aggregated field productivity of eight CS sub-fields over the last 10 years

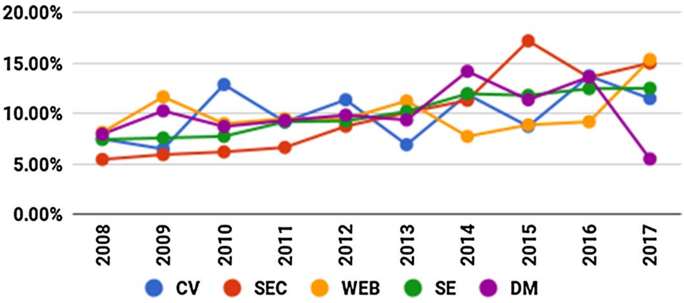


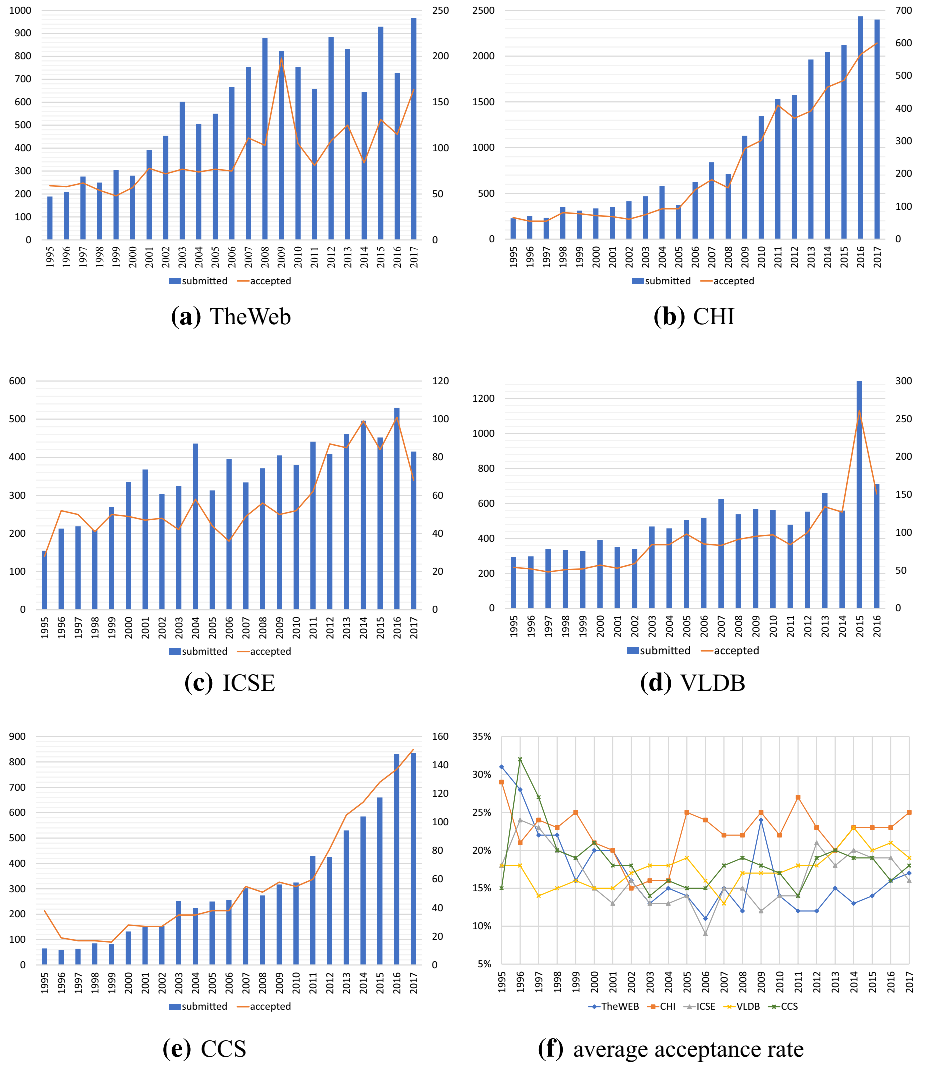
Fig. 9 Field productivity of five CS sub-fields over the last 30 years

metric to the past ten years, because not all data were available for all events in the earlier years. It is observed that CV is the most productive community over the other communities with an FP of 22%, then the DM community comes, while the computational learning community (LRN) is the lowest one of only 4%. As shown in Fig. 9, DM and WEB remained at the same FP with some ups and downs from 2008 to 2013; then, WEB had a slight decline in the next year, then began to rise again until it reached its maximum value in 2017. In 2015, the FP of SEC was the highest among all the others, i.e., about 17%, then dramatically decreased to 13.5% in the next year, then saw a slight increase to 14.9% in 2017. In summary, the FP of all communities has continued to increase gradually since 2008, ranging between 5.5% and 17% in the whole period, with the highest FP ever (17%) for SEC in 2015. In particular, FP of SE varied between 7.4% and 12.5%; for, CV it varied between 6.5% and 13.7%.

Individual events analysis

This section presents a study of the most renowned events within each CS community.

Submissions and accepted papers Fig. 10a–e displays the number of submissions as well as the number of accepted papers of the top-events, i.e., events with the highest h5-index,



(a) TheWeb

(b) CHI

(c) ICSE

(d) VLDB

(e) CCS

(f) average acceptance rate

Fig. 10 Number of submissions (main axis) and accepted papers (secondary axis) per year of the top event in each CS community for the period 1995–2017

in each CS community over the period 1995–2017. Among all events studied, in 2017, CHI had the highest number of submissions (2,400 submissions), while ICSE had the lowest one of 415 submissions. Accordingly, CHI had the highest (600 papers) and ICSE the lowest number of accepted papers (68 papers). Overall, the top events in each CS community show an increasing trend in both submissions and accepted papers (see Table 4).

Average acceptance rate Fig. 11 shows the average acceptance rate (AAR) of each of the top-5 events in 2017 in each CS community along with the country where most editions were hosted. The Web Conference, UIST (ACM Symposium on User Interface Software and Technology), VLDB (International Conference on Very Large Databases), ICSE

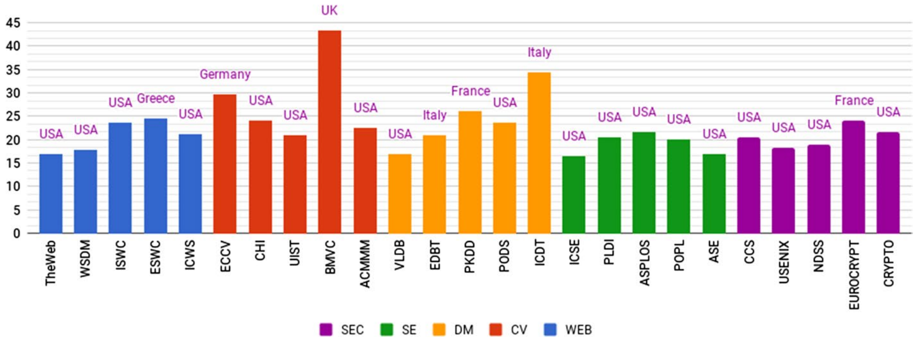


Fig. 11 The average acceptance rate of top-5 events in 2017 in each CS community indicating the most visited country for each event series

(International Conference on Software Engineering), and USENIX (Usenix Security Symposium) have the lowest AAR among the top-5 events within WEB, CV, DM, SE, and SEC respectively. In general, the AAR for the top event in each CS community is in the 13–25% range in the 20 year time window (Fig. 10f). In addition, the acceptance rate of all events has remained fairly stable during the whole period. As can be seen from the charts Fig. 10a–e, the number of submissions has continuously increased over the whole period with slight ups and downs. However, the number of accepted papers increased steadily from the beginning until the end of the period, except for The Web Conference and VLDB in 2009 and 2015, respectively, where they showed peaks. The highest AAR ever, among these events, was the one of CCS in 1996 (32%), which subsequently decreased to 18% in 2017. The AAR of The Web Conference was relatively high (31%) in 1996, then began to decrease until it reached 17% in 2017. The AAR of ICSE dramatically decreased from 24% in 1996 to only 9% in 2006, then increased to 15% in the next year and slightly increased to 16% in 2017. A reason for decreasing acceptance rates is the increasing number of submissions, with the number of presentation slots at an event being more or less constant over time.

Continuity The continuity of TheWeb and ISWC (International Semantic Web Conference) is 100%, whilst they were held every year since their inception. On the other hand, the continuity of CHI (Conference on Human Factors in Computing Systems) is 97% because it was held every year since 1982 except for 1984. We observed that some events, such as ASPLOS and EDBT, have changed the regularity from $R = 2$ to $R = 1$ due to the high demand of submissions. Therefore, we computed the average of the continuity within each of these periods. For instance, EDBT (International Conference on Extending Database Technology) had a regularity of 2 in the period 1988–2008 and then it continued to convene every year. Overall, we observed a very high continuity among the most renowned events (see Table 5).

Progress ratio We calculated the PR of the top-events in each CS community in the period 1997–2016. As shown in Fig. 12, the PR of the top-5 events had a slight rise in the period 1997–2005; then, they all rose noticeably in the last decade. Overall, events of all CS communities have shown a drastic increase in PR since the beginning, particularly, CCS and CHI.

H5-index Fig. 13 shows the top-5 events in five CS communities according to their h5-index, calculated in 2016. The conference with the highest h5-index among all the

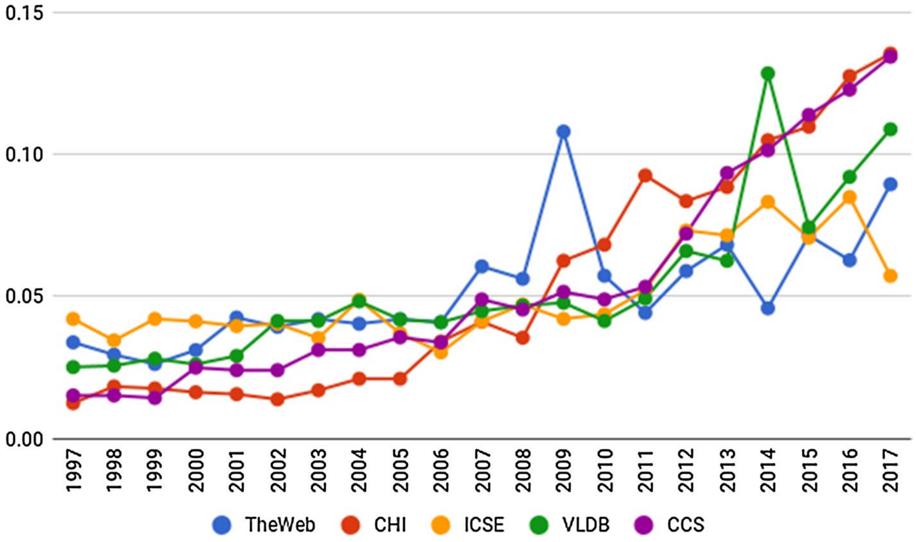


Fig. 12 PR of the top events in each CS sub-community in the last two decades

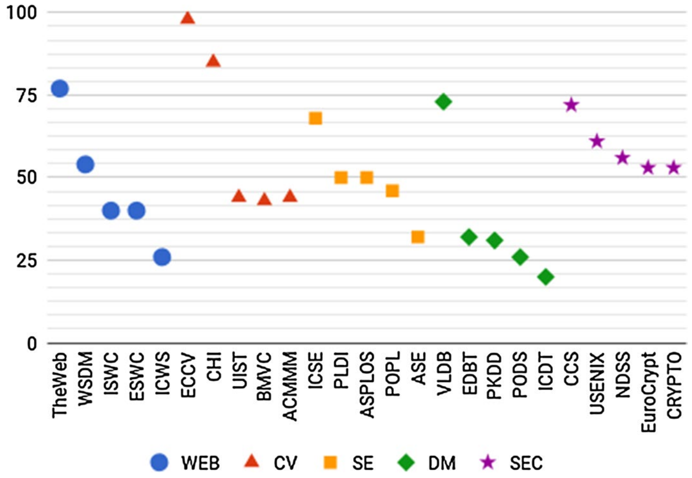


Fig. 13 Top-5 events in each CS community according to their h5-index in 2017

fields is ECCV (European Conference on Computer Vision) with 98 (in the CV field), and TheWeb comes next with 77 (in the Web technologies field). Overall, we observed that the renowned events in CS usually have an h5-index greater than 20.

Geographical distribution For country-level GD, VLDB and TheWeb have $\bar{x} = 1$, which means that they moved to a different country each year, while SP (IEEE Symposium on Security and Privacy) and NDSS have $\bar{x} = 0$, which means that they stayed in the same country every year. For continent-level GD, ESWC (European/Extended Semantic Web Conference) and NDSS were always held in Europe and North America,

Table 5 Continuity of top-5 events in five CS communities

Event	Field	Birth	Age	Editions	Regularity	Continuity (%)
TheWeb	WEB	1995	24	24	1	100
WSDM	WEB	2008	11	11	1	100
ISWC	WEB	2002	17	17	1	100
ESWC	WEB	2004	15	15	1	100
ICWS	WEB	1995	24	24	1	100
ECCV	CV	1992	27	14	2	100
CHI	CV	1982	37	36	1	97
UIST	CV	1988	31	31	1	100
BMVC	CV	1987	32	32	1	100
ACMMM	CV	1993	26	26	1	100
ICSE	SE	1995	24	24	1	100
PLDI	SE	1987	32	32	1	100
ASPLOS	SE	1982	37	23	1	98
POPL	SE	1973	46	45	1	98
ASE	SE	1991	28	28	1	100
VLDB	DM	1985	34	35	1	100
EDBT	DM	1988	31	21	2	100
PKDD	DM	1997	22	22	1	100
PODS	DM	1982	37	37	1	100
ICDT	DM	1986	33	21	2	100
CCS	SEC	1993	26	26	1	100
USENIX	SEC	1990	29	28	1	97
NDSS	SEC	1993	26	25	1	96
EUROCRYPT	SEC	1982	37	37	1	100
CRYPTO	SEC	1995	24	24	1	100

The regularity column shows the most recent regularity of each event

respectively, while ICDE (International Conference on Data Engineering) alternatively moved across continents, i.e., North America, Europe, and Asia.

Time distribution We computed the frequency of occurrence of top-5 events (identified using the SER ranking) for each event each month of the year since its establishment. Table 6 shows the most frequent month in which events take place along with the percentage of occurrence in this month. We observed that most of the renowned events usually take place around the same month each year with a slight shift of maximum one month. For instance, 50% of the editions of TheWeb were held in May and 41% in April. The CVPR conference has been held 28 times (out of 31) in June, and the PLDI conference has been held 33 times (out of 36) in June.This helps potential authors to expect when the event will take place next year and thus helps them with the submission schedule organization.

Table 6 Scientometric profile of the top-5 events in each CS community

Acronym	h5	CORE	Q	TD	AAR (%)	FP (%)	PR (%)	N	GD
TheWeb	77	A*	A1	May (50%)	**17**	6.16	8	23	USA (22%)
WSDM	54	A*	B1	Feb.(91%)	18		14	11	USA (55%)
ISWC	40	A	A1	Oct.(63%)	24		10	21	USA (57%)
ESWC	40	A	A1	May (60%)	24		11	15	Greece (60%)
ICWS	26	A	A1	Jun(35%)	19		17	24	USA (96%)
ECCV	**98**	A	A1	May(33%)	30	5.25	**19**	14	Germany(21%)
CHI	85	A*	A1	Apr.(64%)	22		10	35	USA (57%)
UIST	44	A	A1	Oct.(58%)	21		9	31	USA (68%)
BMVC	43	–	A2	Sep.(89%)	41		6	32	UK (97%)
ACMMM	44	A*	A	Oct.(60%)	22		11	25	USA (52%)
ICSE	68	A*	A1	May (60%)	17	8.38	5	23	USA (43%)
PLDI	50	A*	A1	Jun.(92%)	21		5	31	USA (68%)
ASPLOS	50	A*	A1	Mar.(43%)	22		8	23	USA (87%)
POPL	46	A*	A1	Jan.(89%)	17		11	**44**	USA (80%)
ASE	31	A	A1	Sep.52(%)	21		6	27	USA (59%)
VLDB	73	A*	A1	Aug.59(%)	17	8.35	4	33	USA (12%)
EDBT	32	A	A2	Mar.(100%)	21		4	21	Italy (19%)
PKDD	31	A	A2	Sep.(86%)	26		9	22	France (14%)
PODS	26	A*	A1	Jun.(45%)	24		5	36	USA (81%)
ICDT	**20**	A	B1	Mar.(100%)	34		5	20	Italy (20%)
CCS	72	A*	A1	Oct.(46%)	20	9.23	11	25	USA (72%)
USENIX	61	A*	A1	Aug.(61%)	19		10	28	USA (82%)
NDSS	56	A	A1	Feb.(96%)	19		12	25	USA (100%)
EuroCrypt	53	A*	A1	May (62%)	23		6	36	France (11%)
CRYPTO	53	A*	A1	Aug.(100%)	21		7	23	USA (100%)

N is the total number of editions. Prominent values are bold

Discussion and future work

We analyzed metadata of scholarly events of four scientific fields (Computer Science, Physics, Engineering, and Mathematics) involving conferences, workshops, meetings, and symposiums. We report the results of our analysis of events metadata within the two datasets in the last 30 years according to the proposed metrics suite. The results we obtained from this study have been categorized into three categories: scientific fields analysis, CS sub-fields analysis, and individual events analysis. These results reveal that the long continuity of events highlights the importance of such events for the CS, MATH, PHY, and ENG communities (RQ1). Furthermore, the increasing number of submissions and the growing progress ratio of CS events provide clear evidence of the weight of scholarly events in different CS communities. Researchers consider scholarly events as a serious gate to disseminate their research results. They consider certain characteristics to select the target venue. As a result of domain conceptualization to provide the foundation for this study, a comprehensive list of event-related properties provides empiric evidence on what makes an event high-ranked in its community (RQ2). The results also shed light on the publication policies of researchers in CS, PHY, ENG, and MATH in terms of publication venue and citation count. In the last decades, we observed an increasing trend in both submissions and accepted papers in all CS events. For instance, the average number of submissions, i.e., submissions to renowned events per year, to SEC events has doubled, i.e., from around 900 submissions in the period 2003–2007 to around 1,900 in 2013–2017. We summarize the contributions of this research as follows:

- The creation of a dataset of scholarly events belonging to four scientific fields, which we imported into the scholarly event knowledge graph of OpenResearch.org,
- The conceptualization of the scholarly communication domain and the development of an event quality framework,
- A metrics suite based on the domain conceptualization, which contains new metrics for scholarly events' impact assessment such as continuity, community popularity, field productivity, and progress ratio,
- An empirical evaluation of the quality of scholarly event metadata of CS, PHY, ENG, and MATH research communities involving different event types such as conferences, workshops, meetings, and symposiums,
- A workflow model for data curation and metadata analysis of scholarly events, and
- Support for communities by giving recommendations to different stakeholders of their events.

Generally, the acceptance rate is considered one of the most important characteristics of scholarly events, however, the findings of this study indicate that the success of events depends on several other characteristics as well, such as continuity, the popularity of events' topics, and citations of published papers (reflected by the h5-index of the event). SEQA can be used in further events' metadata analysis and for multi-criteria events ranking (RQ3). After integrating and analyzing the results we obtained from this study, we found that the most noteworthy findings to record are:

- There is not much historic and aggregated information available about publications in the PHY, ENG, and MATH fields,
- Most editions of the renowned events in all communities have been held around the same time of the year with similar deadlines,
- Most of those events that are high-ranked and have a high h5-index also have a long continuity (greater than 90%),
- Among all countries hosting events, the USA has been the host of about 50% of the scholarly events in all communities in the last two decades,
- The field productivity of all CS communities continuously increased since 2008, and the top-events kept the trend of acceptance rates mostly stable over time regardless of the number of submissions.
- Based on the SJR indicator, the CS and MATH communities are more prolific and their publications have more citations among each other, compared to PHY and ENG,
- The CV community had the highest number of submissions and accepted papers during the three 5-year time windows,
- The CS community has the largest number of events with h5-index exceeding 30 compared to other communities, which can be attributed to scholarly events having an even more important role in CS, and

- Most of the research findings of non-CS communities were published as abstracts or posters, while research findings of CS were published as full research articles in formal proceedings.

Based on these findings, a set of recommendations has been concluded to different stakeholders, involving event organizers, potential authors, and sponsors. *Organizers*: Organizers can, by adjusting the topics covered by their events to the most productive and popular ones, increase the impact and the productivity of their events. The possibility of having a progress ratio overview of other events enables organizers to compare their event with competing events and to identify organizational problems, e.g., publicity issues, the reputation of the members, and location dynamicity. Therefore, in order to provide a high-profile event to the community, following certain strategies to comply with the characteristics of high-ranked events is necessary, e.g., keeping event topic coverage up to date with new research trends, involving high-profile people and sponsors, maintain a high continuity of the event, increasing the geographic distribution of event venues, and minimizing the time distribution. *Potential authors*: Community productivity and popularity change the research direction of individual scientists. Submitting to events with a broad range of up-to-date topics keeps the research productivity and publication profile of researchers aligned with growing communities. While searching for a venue to submit research results, considering characteristics of renowned events may influence the future visibility and impact of the submissions, if accepted. *Sponsors and proceedings publishers*: The progress ratio of renowned events and considered characteristics gives insights about events of small size or preliminary events. Sponsoring such small scale, but reliable and valuable events may support their rapid growth and may influence the popularity and overall direction of the associated research topics. Overall, this study helps to shed light on the evolving and different publishing practices in various communities and helps to identify novel ways for scholarly communication, such as the blurring of journals and conferences or open-access overlay-journals as they already started to emerge. In addition, we anticipate that the findings will encourage researchers in MATH, ENG and PHY to publish and archive more information about their events, which will help in the events metadata analysis.

This research is in the process of extension to other fields of science. More metrics are planned to be considered, such as sponsorship and co-authorship analysis and paper affiliation analysis as well as awards. It is also interesting to assess the impact of digitization with regard to journals (which receive more attention than events in fields other than Computer Science). Although large parts of our analysis methodology have already been automated, we are currently in the process of enabling instant, interactive queries to our dataset using a dedicated service, i.e., SPARQL-AG (Fathalla et al. 2019b) In addition, the proposed metrics will be used in providing new and innovative venue rankings for different research fields, thus allowing in particular younger researchers without a long-term experience to identify better publication strategies and consequently contribute more productively to the advancement of research.

Acknowledgements Open Access funding provided by Projekt DEAL. Part of this work has been funded by the DFG project "ConfIDent – A reliable platform for scientific events" (Grant agreements LA 3745/4-1 and SE 1827/16-1) as well as ERC ScienceGRAPH (Grant Agreement ID 819536).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Agarwal, S., Mittal, N., & Sureka, A. (2016). A scientometric analysis of 9 ACM SIGWEB cooperating conferences. *ACM SIGWEB Newsletter(Autumn)*, *6*, 1–15.

Alexiou, G., Vahdati, S., Lange, C., Papastefanatos, G., & Lohmann, S. (2016). Openaire lod services: Scholarly communication data as linked data. In *International workshop on semantic, analytics, visualization* (pp. 45–50).

Ameloot, T. J., Marx, M., Martens, W., Neven, F., & van Wees, J. (2011). 30 years of PODS in facts and figures. *SIGMOD Record*, *40*(3), 54–60.

Aumüller, D., & Rahm, E. (2011). Affiliation analysis of database publications. *SIGMOD Record*, *40*(1), 26–31.

Bakare, V., & Lewison, G. (2017). Country over-citation ratios. *Scientometrics*, *113*(2), 1199–1207.

Barbosa, S. D. J., Silveira, M. S., & Gasparini, I. (2017). What publications metadata tell us about the evolution of a scientific community: The case of the Brazilian human-computer interaction conference series. *Scientometrics*, *110*(1), 275–300.

Behrens, J. T., & Yu, C.-H. (2003). Exploratory data analysis. Handbook of psychology.

Biryukov, M., & Dong, C. (2010). Analysis of computer science communities based on DBLP. In *International conference on theory and practice of digital libraries (tpdl)* (pp. 228–235). Springer.

El-Din, H. M. A., Eldin, A. S., & Hanora, A. M. (2016). Bibliometric analysis of Egyptian publications on Hepatitis C virus from PubMed using data mining of an in-house developed database (HCVDBegy). *Scientometrics*, *108*(2), 895–915.

Fathalla, S. (2018). EVENTS: A dataset on the history of topprestigious events in five computer science communities. In A. González- Beltrán, F. Osborne, S. Peroni, & S. Vahdati (Eds.), *Semantics, analytics, visualization* (pp. 110–120). Cham: Springer.

Fathalla, S., & Lange, C. (2018). EVENTSKG: a knowledge graph representation for top-prestigious computer science events metadata. In *International conference on computational collective intelligence* (pp. 53–63). Springer.

Fathalla, S., Lange, C., & Auer, S. (2019a). EVENTSKG: A 5-star dataset of topranked events in eight computer science communities. In *European semantic web conference* (pp. 427–442).

Fathalla, S., Lange, C., & Auer, S. (2019b). A human-friendly query generation frontend for a scientific events knowledge graph. In *International conference on theory and practice of digital libraries* (pp. 200–214).

Fathalla, S., Vahdati, S., Auer, S., & Lange, C. (2018). Metadata analysis of scholarly events of computer science, physics, engineering, and mathematics. In *International conference on theory and practice of digital libraries* (pp. 116–128).

Fathalla, S., Vahdati, S., Lange, C., & Auer, S. (2017). Analysing scholarly communication metadata of computer science events. In *International conference on theory and practice of digital libraries (tpdl)* (pp. 342–354). Springer.

Fathalla, S., Vahdati, S., Lange, C., & Auer, S. (2019). SEO: A scientific events data model. In *International semantic web conference* (pp. 79–95).

Glass, G. V., Smith, M. L., & McGaw, B. (1981). *Meta-analysis in social research*. Thousand Oaks: Sage Publications. Incorporated.

González-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics*, *4*(3), 379–391.

Guilera, G., Barrios, M., & Gómez-Benito, J. (2013). Meta-analysis in psychology: A bibliometric study. *Scientometrics*, *94*(3), 943–954.

Hedges, L. V. (1984). Advances in statistical methods for meta-analysis. *New Directions for Evaluation*, *1984*(24), 25–42.

Hiemstra, D., Hauff, C., De Jong, F., & Kraaij,W. (2007). SIGIR's 30th anniversary: an analysis of trends in IR research and the topology of its community. In *Acm sigir forum* (Vol. 41, pp. 18–24). ACM.

Jeong, S., & Kim, H.-G. (2010). Intellectual structure of biomedical informatics reflected in scholarly events. *Scientometrics*, *85*(2), 541–551. https://doi.org/10.1007/s11192-010-0166-z.

Martinez, W. L., Martinez, A. R., Martinez, A., & Solka, J. (2010). *Exploratory data analysis with MATLAB*. Boca Raton: CRC Press.

Nascimento, M. A., Sander, J., & Pound, J. (2003). Analysis of SIGMOD's coauthorship graph. *ACM Sigmod Record*, *32*(3), 8–10.

Proceedings of the 4th workshop on semantic publishing (sepublica) (No. 1155). (2014).

Sabharwal, A. (2015). Digital curation in the digital humanities: Preserving and promoting archival and special collections. Chandos Publishing.

Scimago, G. (2007). SJR-SCImago journal & country rank.

Semantic web evaluation challenges. (2014).

Vahdati, S., Arndt, N., Auer, S., & Lange, C. (2016). OpenResearch: Collaborative management of scholarly communication metadata. In *20th international knowledge engineering and knowledge management conference (ekaw)* (pp. 778–793). Springer.

Wilsdon, J. (2016). The road to ref 2021: why i welcome lord stern's blueprint for research assessment. The Guardian. Retrieved June, 13, 2017.

Wolf, F. M. (1986). Meta-analysis: Quantitative methods for research synthesis (Vol. 59). Sage.

Yan, S., & Lee, D. (2007). Toward alternative measures for ranking venues: a case of database research community. In *7th acm/ieee-cs joint conference on digital libraries (jcdl)* (p. 235–244). ACM.

Affiliations

Said Fathalla[1,2] · Sahar Vahdati[3] · Christoph Lange[4,5] · Sören Auer[6,7]

Sahar Vahdati
sahar.vahdati@cs.ox.ac.uk

Christoph Lange
lange@cs.rwth-aachen.de

Sören Auer
auer@tib.eu

[1] Smart Data Analytics (SDA), University of Bonn, Bonn, Germany

[2] Faculty of Science, University of Alexandria, Alexandria, Egypt

[3] Department of Computer Science, University of Oxford, Oxford, UK

[4] RWTH Aachen University, Aachen, Germany

[5] Fraunhofer FIT, Sankt Augustin, Germany

[6] L3S Research Center, Leibniz University of Hannover, Hannover, Germany

[7] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany