EVENTSKG: A Knowledge Graph Representation for Top-Prestigious Computer Science Events Metadata

Said Fathalla[1,2] and Christoph Lange[1,3]

[1] Smart Data Analytics (SDA), University of Bonn, Germany
{fathalla,langec}@cs.uni-bonn.de
[2] Faculty of Science, University of Alexandria, Egypt
[3] Fraunhofer IAIS, Germany

Abstract Digitization has made the preparation of manuscripts as well as the organization of scientific events considerably easier and efficient. In addition, data about scientific events is increasingly published on the Web, albeit often as raw dumps in unstructured formats, immolating its semantics and relationships to other data and thus restricting the reusability of the data for, e.g., subsequent analyses. Therefore, there is a great demand to represent this data in a semantic representation using Semantic Web technologies. In this paper, we present the EVENTSKG dataset to offer a comprehensive semantic descriptions of scientific events of six computer science communities for 40 top-prestigious event series over the last five decades. We created a new, publicly available and improved release of the EVENTSKG dataset as a unified knowledge graph based on our Scientific Events Ontology (SEO). It is of primary interest to event organizers, as it helps them to assess the progress of their event over time and compare it to competing events. Furthermore, it helps potential authors looking for venues to publish their work. We shed light on these events by analyzing the EVENTSKG data.

Keywords: Scientific Events Dataset, Scholarly Communication, Linked Data, Semantic Web, Metadata Analysis, Knowledge Graph

1 Introduction

The exponential growth of Web data places an excessive pressure on researchers who are working on scholarly communication to assess, analyze, and organize this huge amount of data produced every day [12]. This paper introduces the EVENTSKG dataset, a new release of our previously presented EVENTS dataset [7], in the form of knowledge graph (KG), containing 60% additional event series belonging to six CS communities. A notable feature of the new release is the use of our Scientific Events Ontology (SEO)[4] as a reference ontology for event metadata modeling and connect related data that was not previously, i.e. in EVENTS, linked. EVENTSKG is a knowledge graph containing metadata of top-40 prestigious events series as a *unified graph* rather than individual RDF dumps for each event series in the previous release. The benefits of publishing EVENTSKG as linked data are:

[4] http://sda.tech/SEOontology/Documentation/SEO.html

- *Data linking*: establish links between dataset elements so that machines can explore related information,
- *Semantic querying*: Linked Data can be queried using the SPARQL query language,
- *Data enrichment*: inference engines could be used to infer implicit knowledge which does not explicitly exist,
- *Data validation*: semantically validate data against inconsistencies.

Events are linked by research fields, hosting country, and publishers. For instance, EVENTSKG is able to answer competency questions such as:

- What are the events related to "*Computational Intelligence*" with an acceptance rate less than 20% and proceedings published by "*Springer*"?,
- Which countries have hosted most of the events related to "*Semantic web*" over the last 20 years?
- Which of the six CS communities has attracted growing interest (in terms of the number of submissions) in the last 10 years?
- Which of the six CS communities has a growing production (in terms of the number of accepted papers) in the last 10 years?

The main goal of EVENTSKG is to facilitate the analysis of events metadata, by enabling them to be queried using semantic query languages such as SPARQL. A key research question that motivates our work is: *What is the effect of digitization on scholarly communication in computer science events*? In particular, we address the following questions: *a*) *What is the orientation of submissions and corresponding acceptance rates of prestigious events in computer science*? *b*) *How did the number of publications of a CS sub-community fluctuate*? *c*) *Did the date of prestigious events changes from year to year*? *d*) *Which countries host most events in different CS communities*? In terms of events' impact, we address the following questions: *a*) *What are the high-impact events of computer science*? *b*) *How are the high-impact events currently ranked in the available ranking services*? By analyzing the dataset content, we gain some insights to answer these questions. Exploratory data analysis is performed aiming at exploring some facts and figures about CS events over the last five decades. Top-40 prestigious event series have been identified based on several criteria (see subsection 5.1). These event series fall into six CS communities[5]: information systems (IS), security and privacy (SEC), artificial intelligence (AI), computer systems organization (CSO), software and its engineering (SE) and web (WWW).

We believe that EVENTSKG dataset closes an important gap in analyzing the progress of a particular event series and CS community, using prestigious event series in the community, in terms of submissions and publications over a long-term period. Furthermore, it will have a momentous influence on the research community, in particular:

a) *event chairs* – to asses the progress/impact of the event,
b) *potential authors* – to find out events with high impact to submit their work,
c) *proceedings publishers* – to trace the impact of their events.

The rest of the paper is structured as follows: Section 2 presents a brief review of the related work. Section 3 introduces the SEO ontology. Section 4 presents the main characteristics of EVENTSKG. Section 5 explains the curation process of creating and

[5] Using ACM Computing Classification System: `https://dl.acm.org/ccs/ccs.cfm`

evolving the dataset. Section 6 discusses the results of analyzing the EVENTSKG data. Section 7 concludes and outlines our future work.

2 Related Work

Scholarly communication and data publishing have received much attention in the past decade [13, 2, 11, 7, 8, 9]. The first considerable work to provide a comprehensive semantic description of scientific events metadata is the Semantic Web Conference (SWC) ontology [14]. The Semantic Web Dog Food (SWDF) dataset and its successor *ScholarlyData* are among the pioneers of datasets of comprehensive scholarly communication metadata [13]. A first attempt to create a dataset of prestigious events in five computer science communities is represented by our own EVENTS dataset [7]. It covers historical information about 25 top-prestigious events, describing each of them with 15 metadata attributes. The main shortcoming of this dataset is that it published as individual RDF dumps rather than one knowledge graph, by which it loses the potential links between dataset elements. Biryukov and Dong [4] analyzed a sets of top-ranked conferences in different Computer Science Communities and compared them in terms of publication growth rate, population stability and collaboration trends using DBLP. In a different work, we analyzed the evolution of key characteristics of CS conferences over a period of 30 years, including frequency, geographic distribution, and submission and acceptance numbers [8]. Hiemstra et al. [11] analyzed the SIGIR community in terms of authors' countries, number of papers per year for each country and co-authorship. Yan and Lee [18] proposed two measures for ranking academic venues by defining the goodness of a venue. Vasilescu et al. [17] presented a dataset of eleven prestigious software engineering conferences, such as ICSE and ASE, containing accepted papers along with their authors, programme committee members and the number of submissions each year. Agarwal et al. [1] analyzed the bibliometric metadata of seven ACM conferences in information retrieval, data mining, and digital libraries.

Despite these continuous efforts, none of these previous works was done on top of a unified KG. What additionally distinguishes our work from the related work mentioned above is that our analysis is based on a comprehensive list of metrics, considering quality in terms of event-related metadata in six CS communities and the dataset is published as a unified knowledge graph of all events.

3 Scientific Events Ontology

It is considered a good practice to reuse vocabularies from well-known ontologies wherever possible in order to facilitate ontology development and to lower the barrier for third party ontology-aware services reuse one's Linked Data. In this section, we introduce the scientific events ontology (SEO), a reference ontology for modeling data about scientific events such as conferences, symposiums, and workshops. SEO reuses several well-designed ontologies, such as SWC[6], FOAF, SIOC, Dublin Core and

[6] `http://data.semanticweb.org/ns/swc/swc_2009-05-09.html`

SWRC (Semantic Web for Research Communities), and defines some of its own vocabulary as discussed later in this section. The vocabulary of SEO is defined in a new namespace prefixed *seo*, e.g., as seo:EventTrack and seo:Symposium. All namespace prefixes are used according to prefix.cc[7]. Several classes have been used such as ConferenceSeries, AcademicEvent and NonAcademicEvent and data properties such as acronym, startDate and endDate, and object properties such as hasLocation, hasTopic, and isProceedingsOf. Namely, Topic for describing a research topic or scientific area of an event (e.g., Database systems), OrganisedEvent for describing events related to an academic or non-academic event and Role for describing the roles held by people involved in an event. We define inverse relations, e.g., as seo:isTrackOf is the inverse relation of seo:hasTrack and seo:isSponsorOf is the inverse relation of seo:hasSponsor. Thus, if an event E seo:hasTrack T, then it can be inferred that T seo:isTrackOf E. Also, some symmetric relations are defined, such as seo:colocatedWith. Such definitions allow to reveal implicit information and increase the coherence and thus the value of event metadata [10].

4 Characteristics of the EVENTSKG Dataset

EVENTSKG covers three types of events since 1969: conferences, workshops, and symposia[8]. It contains metadata of 1048 editions of 40 events series with 15 attributes each. It is available in four different formats: RDF/XML, Turtle, CSV, and JSON-LD. The number of submissions and publications of each event involves all tracks' submissions and publications. There are several challenges to pursuing the maintenance of EVENTSKG for the future and keeping it sustainable; here is how we address them:

- *Availability:* EVENTSKG is publicly available online under a persistent URL (PURL): `http://purl.org/events_ds`. It is subjected to the Creative Commons Attribution license.
- *Extensibility:* There are three dimensions to extend the dataset to meet future requirements: a) increase the number of events in each community, b) cover more CS communities and c) add more event properties such as hosting organization, registration fees and event sponsors.
- *Validation:* we perform two types of validation: *syntactic* and *semantic* validation. We syntactically validate EVENTSKG to conform with the W3C RDF standards using the online RDF validation service[9] and semantically validate it using Protégé reasoners.
- *Documentation:* the documentation of the dataset has been checked using the W3C Markup Validation Service[10] and is available online on the dataset web page[11].

A concrete use case for querying EVENTSKG is supporting the research community in taking decisions on what event to submit their work to, or whether to accept invitations for being a chair or PC member. For example, finding all events related to

[7] namespace look-up tool for RDF developers: `http://prefix.cc/`

[8] Symposium is a small scale conference, with a smaller number of participants.

[9] `https://www.w3.org/RDF/Validator/`

[10] `https://validator.w3.org/`

[11] `http://sda.tech/EVENTS-Dataset/EVENTS.html`

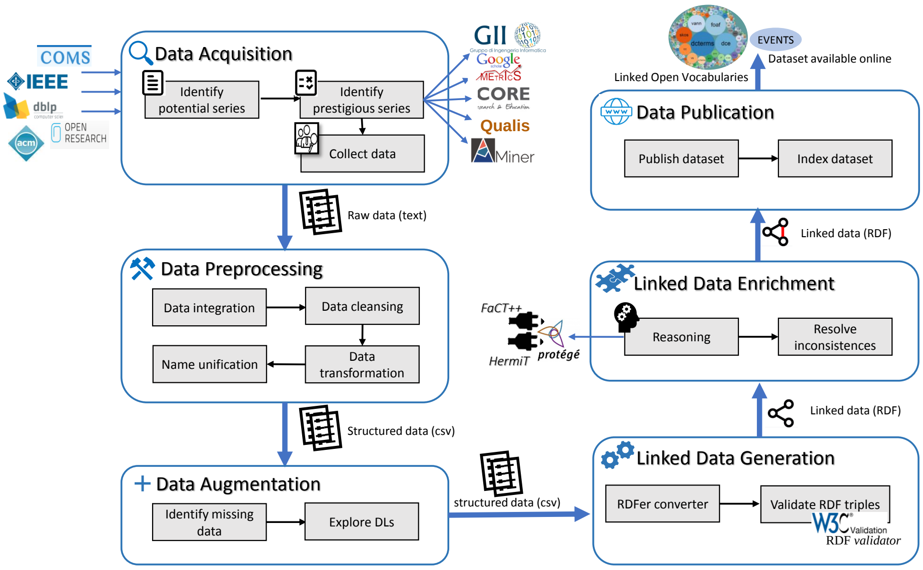

Figure 1: Data curation of EVENTSKG.

Artificial Intelligence, which took place in the USA along with their acceptance rate; this requires joins between Field/Topic and event entities:

```
SELECT ?event ?eventTitle ?acc ?topic WHERE {
  ?event        rdfs:label           ?eventTitle .
  ?event        seo:country          http://dbpedia.org/resource/United_States .
  ?event        seo:field            ?topic
  ?event        seo:acceptanceRate   ?acc .
  ?topic        rdfs:label           "Artificial Intelligence" .}
```

5 Data Curation

While collecting data from different sources, several problems have been encountered, such as data duplication, incomplete data, incorrect data, and the change of event title over time. Therefore, a data curation process has been carried out comprising data acquisition, preprocessing, augmentation, Linked Data generation, data enrichment, and publication. *Data Curation* refers to the activities related to data organization, integration, annotation, and publication collected from various sources [16]. The curation of EVENTSKG dataset is an incremental process involving (see Figure 1): Data acquisition and completion, Data Preprocessing, Data Augmentation, Linked Data Generation, Linked Data Enrichment and Data Publication.

5.1 Data Acquisition and Completion

First, top prestigious events have been identified based on several criteria, such as events ranking services (e.g., CORE, Qualis and GII rankings[12]) and Google h5-index. Second, we have collected the metadata of these events from different data sources, such as IEEE Xplore, ACM DL, DBLP, OpenResearch.org and events' official websites. Data has been collected in either structured or unstructured format to be exposed as Linked Data.

5.2 Data Preprocessing

To prepare the raw data for the further steps, we carried out four preprocessing processes: *data integration*, *data cleansing*, *data structure transformation* and *event name unification*. In the following we briefly describe the main steps of preprocessing the dataset:

- *Data integration:* involves integrating collected data from disparate sources into a unified view.
- *Data cleansing:* Data integration might result in data redundancy, therefore, data cleansing is crucial. This process involves eliminating redundancy, and identifying and mending unsound data.
- *Data structure transformation:* involves transforming cleaned data to a structured format, such as CSV, RDF/XML and JSON-LD.
- *Event name unification:* involves the unification of the names of all editions of an event series which changed their name. The most recent name has been selected because it is the current name of the event.

5.3 Data Augmentation

The objective of the data augmentation process is to add new events to the dataset and fill in missing data. To achieve this objective, we periodically explore online digital libraries for the missing information. The output of this process is structured data in CSV format.

5.4 Linked Data Generation

The adoption of the Linked Data best practices has led to the enrichment of data published on the Web by connecting data from diverse domains such as scholarly communication, people, digital libraries, and medical data [6]. The objective of the Linked Data Generation process here is to generate linked data from unlinked data in the CSV format. We developed RDFer, a custom Java tool to convert data from CSV format to linked data (RDF/XML syntax). Therefore, the next step is to validate (i.e. syntactic validation) the produced data using a standard validation tool (e.g., the W3C RDF online validation service). A sample of the output of RDFer, in RDF/XML syntax, for ICDE in 2017 can be found on the EVENTSKG documentation page.

[12] `http://www.core.edu.au/`, `http://qualis.ic.ufmt.br/`, `https://goo.gl/3kDfFB`

5.5 Linked Data Enrichment

The Linked Data enrichment (LDE) process is important in order to discover the interlinking relationships between RDF triples by using inference engines, i.e., reasoners. The input of LDE is the RDF triples produced by the Linked Data generation and the output is a set of *consistent* RDF triples, including the newly discovered relationships, where available. Semantic inference can be used to improve the quality of data integration in a dataset by discovering new relationships, detecting possible inconsistencies and inferring logical consequences from a set of asserted facts or axioms in an ontology. To enrich and validate (i.e. semantics validation) RDF data, generated from the previous process, we use two reasoners integrated in Protégé, FaCT++ and HermiT[13], which support three types of reasoning: (1) detecting inconsistencies, (2) identifying subsumption relationships, and (3) instance classification [15]. Detecting inconsistencies is a crucial step in LDE because inconsistency results in false semantic understanding and knowledge representation. We resolve detected inconsistencies and run the reasoner again to ensure that no other inconsistencies arise.

5.6 Data Publication

The goal of Linked Data publishing is to enable humans and machines to share structured data on the Web. EVENTSKG is published according to the Linked Data community best practices [3, 5] and registered in a GitHub repository (`https://github.com/saidfathalla/EVENTS-Dataset`). The final step is to index the dataset in a public data portal (e.g., DataHub), which is the fastest way for individuals and teams to find, share and publish high-quality data online. Events is published at DataHub at `https://datahub.ckan.io/dataset/eventskg`.

6 Dataset Content Analysis

Dataset content has been analyzed to answer the research questions presented in section 1. Table 1 provides the results obtained from the preliminary analysis of the dataset[14]. The following metrics are used for the analysis:

Submissions and publications: we see a clear upward trend in the number of submitted and accepted papers during the whole time span, while, roughly speaking, the acceptance rate remains the same.

Time distribution: we observe that prestigious events usually take place around the same month each year (i.e. usual month in Table 1). This helps potential authors to expect when the event will take place next year, which helps for submission schedule organization. *Usual month* refers to the number of times an event has occurred in a specific month. Namely, CVPR conference has been held 26 times (out of 28) in June and POPL has been held 41 times (out of 45) since 1973 in January.

[13] `https://github.com/ethz-asl/libfactplusplus`, `https://github.com/phillord/hermit-reasoner`

[14] This table is an extension to the table on the scientometric profile of events series in the EVENTS dataset as presented by Fathalla et al. [7].

Table 1: Scientometric profile of newly added events to EVENTSKG. N is the number of editions in 2018.

Acronym	Comm.	CORE 2018	GII	Q	h5	N	Avg. AR	Most freq. Country	Usual Month	Usual Month Freq.	Since	Publisher
IJCAR	AI	A*	A	B1	45	10	0.41	UK	Jul	4	2001	ACM
COLT		A*	A+	A2	33	31	0.49	US	Jun	11	1988	PMLR
KR		A*	A+	A2	26	16	0.28	US	Apr	4	1989	AAAI
ISMAR		A*	A	A2	26	21	0.24	US	Oct	10	1999	IEEE
VR		A	A-	A2	17	25	0.26	US	Mar	24	1993	IEEE
FOGA		A*	A-	B3	9	14	0.59	US	Jan	7	1990	ACM
PODS	IS	A*	A+	A1	26	37	0.24	US	Jun	17	1982	ACM
POPL	SE	A*	A++	A1	46	45	0.20	US	Jan	41	1973	ACM
OOPSLA		A*	A++	A1	37	33	0.22	US	Oct	26	1986	ACM
OSDI	SEC	A*	A+	A1	39	13	0.16	US	Oct	7	1994	USENIX
TheWeb	WWW	A*	A++	A1	75	23	0.17	US	May	12	1989	TheWeb
WSDM		A*	A+	B1	54	11	0.18	US	Feb	10	2008	ACM
ISWC		A	A+	A1	40	21	0.24	US	Oct	12	1997	Springer
ESWC		A	A	A1	40	15	0.25	Greece	May	9	2004	Springer
ICWS		A	A	A1	26	25	0.21	US	Jun	6	1995	IEEE

Acceptance rate: Avg. AR of an event series refers to the average of the acceptance rate of all editions. In the last five decades, we observe that the *Avg. AR* for all event series falls between 15% to 30%, except for FOGA, COLT, and IJCAR. Roughly speaking, the largest acceptance rate is the one of FOGA of 59%, while PERCOM has the smallest one of 15%.

H5-index: CVPR has the largest h5-index of 158, while FOGA has the smallest one of 9. Among all considered CS communities, SEC has the largest average h5-index of 58.16, while CSO has the smallest one of 40.2.

Geographical distribution: we analyze the geographical distribution of each event series by recording the country which hosted the most editions of the series (*Most freq. country*). The remarkable observation to emerge from this analysis is that US hosted most editions of events in all communities and most editions of all SE events. France comes second, having hosted most editions of PKDD and EuroCrypt.

Publishers: we observe that several events series organizers publish the proceedings of their events on their own digital library, such as AAAI, VLDB, and TheWeb. On the other hand, ACM publishes the proceedings of most events, and IEEE comes next.

History: IJCAI is the oldest series since it has been established in 1969 (i.e. 50 editions), while RecSys is the most recent one since it has been established in 2007 (i.e. 12 editions).

7 Conclusions and Future work

We present a new release of the EVENTS dataset, called EVENTSKG, as a unified Knowledge graph of top-40 prestigious events based on the SEO ontology. To the best of our knowledge, this is the first time a dataset is published as a knowledge graph of metadata of prestigious events in IS, SEC, AI, CSO, SE, and WWW. EVENTSKG closes an important gap in analyzing the progress of a CS community in terms of submissions and publications and it is of primary interest to steering committees, proceedings publishers and prospective authors. The most striking findings to emerge from analyzing EVENTSKG content is that:

- Among all considered CS communities, SEC has the largest average h5-index, while CSO has the smallest one,
- The number of submissions has kept growing over the last five decades, while, roughly speaking, the acceptance rate has remained the same. The reason may be the digitization of scholarly communication.
- The average acceptance rate for all events, since the first edition, falls into the range 15% to 31%,
- US leads by far, having hosted most editions of events in all communities and most editions of all SE events,
- ACM publishes most of the proceedings of the event, and IEEE comes next.

These findings highlight the usefulness of EVENTSKG for events organizers as well as CS researchers.

To further our research, we are planning to add more events from other CS communities such as computer vision, data management and computational learning and elaborate on the set of features that could be used to efficiently compare events in the same community, such as acceptance rate, h-index, and organizers' reputation, defined, e.g., in terms of their h-index and i10-index. Furthermore, we are planning to perform more complex semantic data analysis by querying EVENTSKG using auto-generated SPARQL queries from user selections using a web service.

Acknowledgments

Said Fathalla would like to acknowledge the Ministry of Higher Education (MoHE) of Egypt for providing a scholarship to conduct this study.

References

1. Agarwal, S., Mittal, N., Sureka, A. A glance at seven acm sigweb series of conferences. In: ACM SIGWEB Newsletter(Summer) (2016), p. 5.
2. Barbosa, S. D. J., Silveira, M. S., Gasparini, I. What publications metadata tell us about the evolution of a scientific community: the case of the Brazilian human–computer interaction conference series. In: Scientometrics 110(1) (2017), pp. 275–300.
3. Berrueta, D., Phipps, J., Miles, A., Baker, T., Swick, R. Best practice recipes for publishing RDF vocabularies. In: Working draft, W3C (2008). `http://www.w3.org/TR/swbp-vocab-pub/`.
4. Biryukov, M., Dong, C. Analysis of computer science communities based on DBLP. In: *TPDL*. Springer. 2010.
5. Bizer, C., Cyganiak, R., Heath, T. et al. How to publish linked data on the web. In: (2007). `http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/`.
6. Bizer, C., Heath, T., Berners-Lee, T. Linked data-the story so far. In: International journal on semantic web and information systems 5(3) (2009), pp. 1–22.

7. Fathalla, S., Lange, C. EVENTS: A Dataset on the History of Top-Prestigious Events in Five Computer Science Communities. In: *International Workshop on Semantic, Analytics, Visualization*. Springer. 2018, In Press.
8. Fathalla, S., Vahdati, S., Lange, C., Auer, S. Analysing Scholarly Communication Metadata of Computer Science Events. In: *International Conference on Theory and Practice of Digital Libraries*. Springer. 2017, pp. 342–354.
9. Fathalla, S., Vahdati, S., Auer, S., Lange, C. Towards a Knowledge Graph Representing Research Findings by Semantifying Survey Articles. In: *International Conference on Theory and Practice of Digital Libraries*. Springer. 2017, pp. 315–327.
10. Gangemi, A. Ontology design patterns for semantic web content. In: *International semantic web conference*. Springer. 2005, pp. 262–276.
11. Hiemstra, D., Hauff, C., De Jong, F., Kraaij, W. SIGIR's 30th anniversary: an analysis of trends in IR research and the topology of its community. In: *ACM SIGIR Forum*. Vol. 41. 2. ACM. 2007, pp. 18–24.
12. McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., Barton, D. Big data: the management revolution. In: Harvard business review 90(10) (2012), pp. 60–68.
13. Möller, K., Heath, T., Handschuh, S., Domingue, J. Recipes for semantic web dog food—The ESWC and ISWC metadata projects. In: *The Semantic Web*. Springer, 2007, pp. 802–815.
14. Nuzzolese, A. G., Gentile, A. L., Presutti, V., Gangemi, A. Semantic web conference ontology-a refactoring solution. In: *International Semantic Web Conference*. Springer. 2016, pp. 84–87.
15. Rubin, D. L., Knublauch, H., Fergerson, R. W., Dameron, O., Musen, M. A. Protege-owl: Creating ontology-driven reasoning applications with the web ontology language. In: *AMIA Annual Symposium Proceedings*. Vol. 2005. American Medical Informatics Association. 2005, p. 1179.
16. Sabharwal, A. Digital curation in the digital humanities: Preserving and promoting archival and special collections. Chandos Publishing, 2015.
17. Vasilescu, B., Serebrenik, A., Mens, T. A historical dataset of software engineering conferences. In: *Proceedings of the 10th Working Conference on Mining Software Repositories*. IEEE Press. 2013, pp. 373–376.
18. Yan, S., Lee, D. Toward alternative measures for ranking venues: a case of database research community. In: *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. ACM. 2007, pp. 235–244.