

CS441 HW2: MapReduce Model for Parallel Processing of DBLP Dataset

I. Background

This project consists of a MapReduce job on the DBLP computer science bibliography dataset in XML format using the Apache Hadoop 3.2.1 framework. The outputs of the job provide information about authors, conferences, journals, venues (e.g., articles, conferences, books, PhD theses, Master's theses), numbers of co-authors, years of publications, and numbers of publications produced at various events and by respective journals. An authorship score is assigned to each author, which is used to rank the top 100 authors and the bottom 100 authors of the entire dataset. Head authors are granted a raise in their scores for their contributions to the publications, whereas tail authors' scores are reduced by the same amount. That is,

1. A **base** score of $1/N$, where N is the number of co-authors, is allocated to each co-author of a publication. If there is only one author for a publication, the author receives the full point.
2. The **head** author of a publication receives a **debit** adjustment of $1/(4N)$ to his or her score.
3. The **last** author of a publication receives a **credit** adjustment of $1/(4N)$ to his or her score.

The full dataset is located at: <https://dblp.uni-trier.de/xml/>. First, the XML dataset is split into smaller subsets, which are sent to an equivalent number of mappers for processing. At this stage, key—value pairs are formed of the Hadoop I/O types **Text** and **FloatArrayWritable**, respectively. The Hadoop types are encoded in the UTF-8 Unicode standard and are utilized by the Hadoop MapReduce framework as they provide built-in methods for serialization and deserialization between map and reduce tasks. This becomes important in terms of efficiency when hundreds of thousands of key—value pairs are emitted through various processes running in parallel. **FloatArrayWritable** is a custom type that extends the **ArrayWritable** class, which encapsulates an array of type **Writable**. In this case, type **FloatWritable** was chosen as it provides adequate precision for holding authorship scores and median and average numbers of co-authors, while maintaining a similar expense in size (e.g., in bytes) as type **IntWritable**.

II. MapReduce

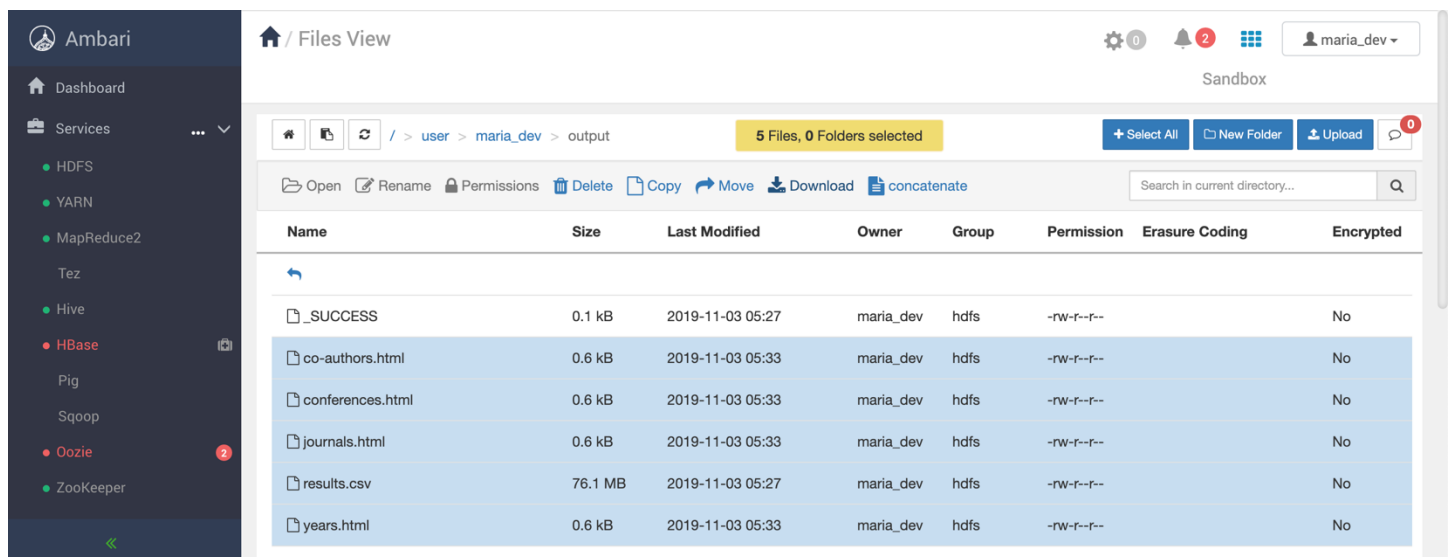
The MapReduce implementation in this project is comprised of 5 phases.

- a. First, the **XmlInputFormat** class, which extends **TextInputFormat**, returns a custom record reader that uses the pre-defined start and end tags (e.g., located in *application.conf*) to scan through the input stream byte-by-byte to return publication records that are sent to mappers as key—value pairs, where keys are positions in the XML input file (e.g., type **LongWritable**), and values are the lines of text that contain the publication record (e.g., type **Text**).
- b. During the **mapper #1** phase, publication records are mapped into various key—value pairs. For each publication, the venue type is ascertained, as well as any authors or editors that are listed in the record. If author(s) is/are listed, their names are stored, through which their count is determined, and each co-author's score is calculated using the formula described above. Then, each author is emitted as a key—value pair, where the key is the author's name, and the value consists of the number of co-authors in the publication and the score the author received for his or her contribution. Similarly, if author(s) is/are listed, the venue is mapped as a key—value pair, where the key is the venue type and the value is the number of co-authors in the publication. If the venue type is ascertained to be a conference or a journal, key—value pairs with value one are emitted, where the keys represent the names of conferences or respective journals. In order to group numbers of co-authors into bins that can be used to plot a histogram that describes publications by co-authors, the number of co-authors in the publication are matched into a corresponding bin, which is mapped as the key in a key—value pair with value one. If the year of the publication is listed, it is matched into its corresponding decade bin, mapped as the key in a key—value pair with value one.
- c. During the **reducer** phase, key—value pairs emitted by mapper #1 are reduced. For author key—value pairs (e.g., where the key represents a single author), authorship score values are summed, and numbers of co-authors are merged into a single set. Similarly, for venue key—value pairs (e.g., where the key represents a single venue type), numbers of co-authors are merged. For all other key—value pairs, counts are summed and updated.

- d. During the **mapper #2** phase, key—value pairs that have already passed through the reducer phase undergo their final mappings. For both author and venue type key—value pairs, sets of numbers of co-authors are sorted and transformed into smaller sets that each contain the total number of publications, the maximum number of co-authors, the median number of co-authors, and the average number of co-authors. Author key—value pairs also retain cumulative authorship scores for each author, who are ranked into lists of the top 100 and the bottom 100 authors. For conference and journal key—value pairs, counts are also placed into bins. All final key—value pairs are emitted.
- e. Finally, the **CsvOutputFormat** class gets the default path for the output with **.csv** extension, sets its base name (e.g., **results**) and returns a custom record writer that writes key—value pairs into the data output stream with a comma separator in accordance with the CSV file format.

III. Results

The results of the MapReduce job can be obtained using Ambari's Files View UI, conveniently accessible at: <http://sandbox-hdp.hoziprtonworks.com:8080/>. From the user's output directory, select **results.csv** and the Plotly charts named **co-authors.html**, **conferences.html**, **journals.html** and **years.html**. Click Download.



Name	Size	Last Modified	Owner	Group	Permission	Erasure Coding	Encrypted
_SUCCESS	0.1 kB	2019-11-03 05:27	maria_dev	hdfs	-rw-r--r--		No
co-authors.html	0.6 kB	2019-11-03 05:33	maria_dev	hdfs	-rw-r--r--		No
conferences.html	0.6 kB	2019-11-03 05:33	maria_dev	hdfs	-rw-r--r--		No
journals.html	0.6 kB	2019-11-03 05:33	maria_dev	hdfs	-rw-r--r--		No
results.csv	76.1 MB	2019-11-03 05:27	maria_dev	hdfs	-rw-r--r--		No
years.html	0.6 kB	2019-11-03 05:33	maria_dev	hdfs	-rw-r--r--		No

The **results.csv** file consists of a series of 9 sets.

```
0,<AUTHOR NAME>,<SCORE>,<NUM PUB>,<MAX>,<MED>,<AVG>
0,Chris Kanich,12.447,46,15,4,4.543
0,Dale Reed,5.369,16,13,4.5,5.188
0,Emanuelle Burton,3.042,9,6,3,3.667
0,G. Elisabeta Marai,11.209,44,9,5,4.773
0,Joe Hummel,1.513,6,7,5.5,5.333
0,Mark Grechanik,23.548,69,9,3,3.739
0,Mitchell D. Theys,5.715,23,11,5,5.609
0,Nasim Mobasher,2.817,10,5,3,3.4
0,Patrick Troy,0.486,2,9,5.5,5.5
```

Set 0 is the entire list of **authors** in the dataset in alphabetical order. It consists of:

- | | |
|---------------------------------|------------------------------|
| (1) author names, | (2) authorship scores, |
| (3) numbers of publications, | (4) maximum # of co-authors, |
| (5) median # of co-authors, and | (6) average # of co-authors. |

```
1,<CONF KEY>,<COUNT>
1,globecom,18948
1,icassp,36344
1,icc,21571
1,icra,22608
1,igarss,21850
1,interspeech,21064

2,<JOURNAL NAME>,<COUNT>
2,Applied Mathematics and Computation,18427
2,CoRR,235628
2,Discrete Mathematics,14310
2,IACR Cryptology ePrint Archive,13893
2,IEEE Access,22552
2,Sensors,20483
```

Set 1 is the entire list of **conferences** in the dataset in alphabetical order. It consists of:

- (1) conference keys, and
- (2) numbers of publications.

Set 2 is the entire list of **journals** in the dataset in alphabetical order. It consists of:

- (1) journal names, and
- (2) numbers of publications.

```
3,<VENUE TYPE>,<NUM PUB>,<MAX>,<MED>,<AVG>
3,articles,2125403,287,3,2.947
3,books,51188,50,2,2.343
3,conferences,2522130,155,3,3.159
3,master's theses,12,1,1,1
3,phD theses,74115,3,1,1.002
```

Set 3 consists of all the types of **venues** in the dataset in alphabetical order. It consists of:

- (1) venue types,
- (2) numbers of publications,
- (3) maximum # of co-authors,
- (4) median # of co-authors, and
- (5) average # of co-authors.

When the program runs, this data tabulates and prints to console for easier viewing in the following format.

```
===== Venues =====
| Venue          | Num Pub | Max | Median | Average |
| articles       | 2125403 | 287 | 3.0    | 2.9     |
| books          | 51188   | 50  | 2.0    | 2.3     |
| conferences    | 2522130 | 155 | 3.0    | 3.2     |
| master's theses | 12      | 1   | 1.0    | 1.0     |
| phD theses     | 74115   | 3   | 1.0    | 1.0     |
```

Sets 4 - 7 consist of bins with numbers of publications for co-authors, years, conferences and journals, respectively. When the program runs, this data tabulates and prints for easier viewing in the following format.

===== Co-authors =====			===== Journals =====	
Num Co-authors	Num Pub		Num Pub	# Journals
1 co-author	812090		1-199	571
2-3 co-authors	2481779		200-499	419
4-6 co-authors	1305908		500-1199	387
7-9 co-authors	137770		1200-2399	220
10+ co-authors	35301		2400+	211

===== Conferences =====		===== Years =====	
Num Pub	Num Conf	Decade	Num Pub
1-199	2792	1970s & earlier	55519
200-599	1098	1980s	128386
600-1199	500	1990s	442477
1200-1999	235	2000s	1403224
2000+	228	2010s	2795755

Last but not least, sets 8 and 9 are lists for the top 100 authors by authorship score in descending order and the bottom 100 authors by authorship score in ascending order. When the program runs, this data tabulates and prints to console for easier viewing in the following format.

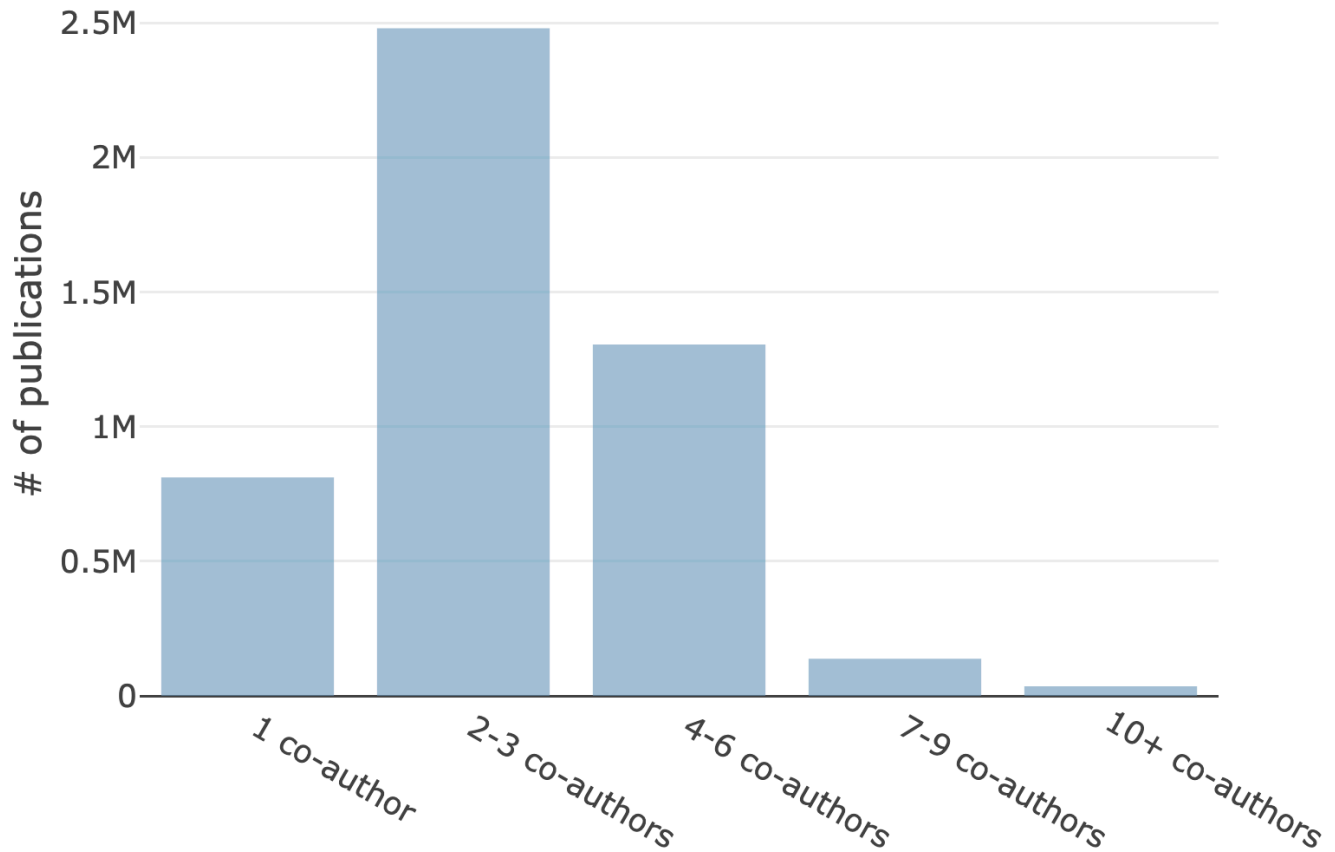
===== Top 100 Authors =====						
Score	Author	Num Pub	Max	Median	Average	
443.007	Ronald R. Yager	624	7	1.0	1.9	
441.554	H. Vincent Poor	1785	13	4.0	3.6	
403.234	Witold Pedrycz	1076	12	3.0	3.1	
395.963	Irith Pomeranz	638	6	2.0	2.2	
371.713	Wei Wang	1310	96	4.0	4.4	
359.400	Wei Li	1226	96	4.0	4.6	
357.400	Wei Zhang	1303	105	4.0	4.5	
354.521	Yu Zhang	1146	25	4.0	4.3	
354.017	T. D. Wilson 0001	373	5	1.0	1.1	
349.862	Mohamed-Slim Alouini	1409	14	4.0	3.6	
349.707	Elisa Bertino	1032	20	3.0	3.5	
344.506	Chin-Chen Chang 0001	914	8	3.0	3.0	
336.005	Xin Wang	1139	17	4.0	4.3	
330.204	Philip S. Yu	1355	14	4.0	4.0	
327.095	Vladik Kreinovich	656	11	2.0	2.6	
319.781	David Eppstein	590	22	2.0	2.8	
314.104	Joseph Y. Halpern	565	6	2.0	2.2	
313.308	Noga Alon	627	9	3.0	3.0	
305.264	Yang Liu	1060	45	4.0	4.5	
304.780	Yong Wang	903	105	4.0	4.1	
...						

Bottom 100 Authors						
Score	Author	Num Pub	Max	Median	Average	
0.004	Yutaka Nakachi	1	264	264.0	264.0	
0.005	Marcel Zoll	2	287	287.0	287.0	
0.006	Omar Zapata	1	118	118.0	118.0	
0.006	Álvaro Iglesias-Arias	1	155	155.0	155.0	
0.007	K. Yates	1	115	115.0	115.0	
0.007	Yolanda Sestayo de la...	2	287	287.0	287.0	
0.007	Zexiong Cai	1	139	139.0	139.0	
0.008	Yong Song Gho	1	95	95.0	95.0	
0.008	Maryam Kavousi	1	92	92.0	92.0	
0.008	V. Reita	1	119	119.0	119.0	
0.008	Zahari Kassabov	1	118	118.0	118.0	
0.009	R. D. Jackson	1	115	115.0	115.0	
0.009	Mark G. Aartsen	2	287	287.0	287.0	
0.009	Valentin Bisson	1	112	112.0	112.0	
0.010	Zhiyao Guo	1	105	105.0	105.0	
0.010	Nianhao Xie	1	104	104.0	104.0	
0.010	Timothy L. Thomas	1	102	102.0	102.0	
0.010	Zhu L. Yang	1	101	101.0	101.0	
0.010	Vishwas Chitale	1	99	99.0	99.0	
0.010	Susan J. Fisher	1	96	96.0	96.0	

...

The *co-authors.html*, *conferences.html*, *journals.html* and *years.html* files consist of graphical representations of the bins in histogram format.

publications by co-authors



As can be observed in this histogram, the great majority of publications listed in the DBLP dataset (approaching 2.5 million) have between 2-3 co-authors. This is followed by publications that have 4-6 co-authors (~1.3 million) and, subsequently, by publications that have a single co-author (~800k).

The *co-authors.html*, *conferences.html*, *journals.html* and *years.html* files consist of graphical representations of the bins in histogram format.

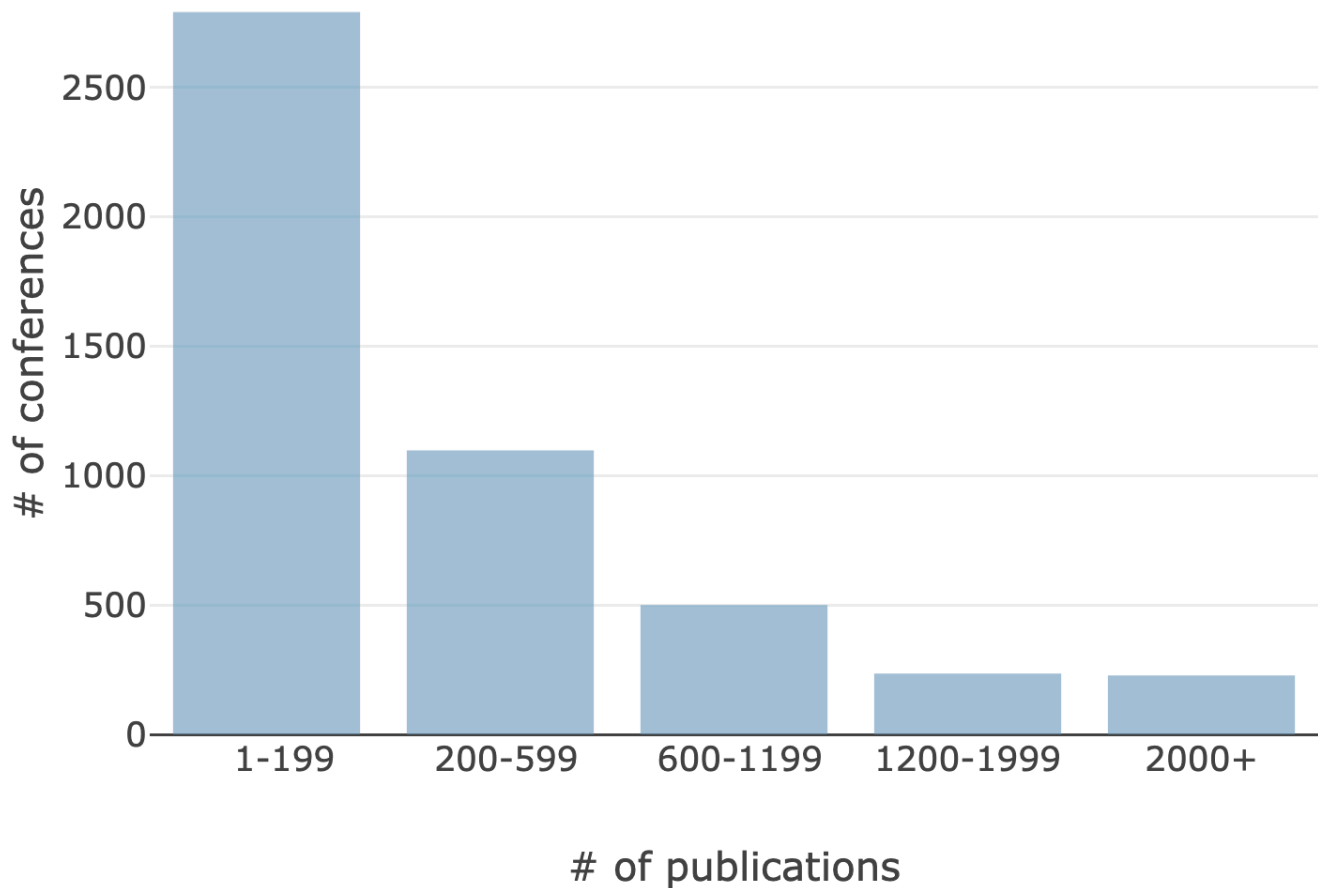
publications by journals



As can be observed in this histogram, the majority of journals have published between 1 and 199 publications. This is followed by journals that have published between 200 and 499 publications and, subsequently, by journals that have published between 500 and 1199 publications. The clear trend is that few journals have published a larger number of publications.

The *co-authors.html*, *conferences.html*, *journals.html* and *years.html* files consist of graphical representations of the bins in histogram format.

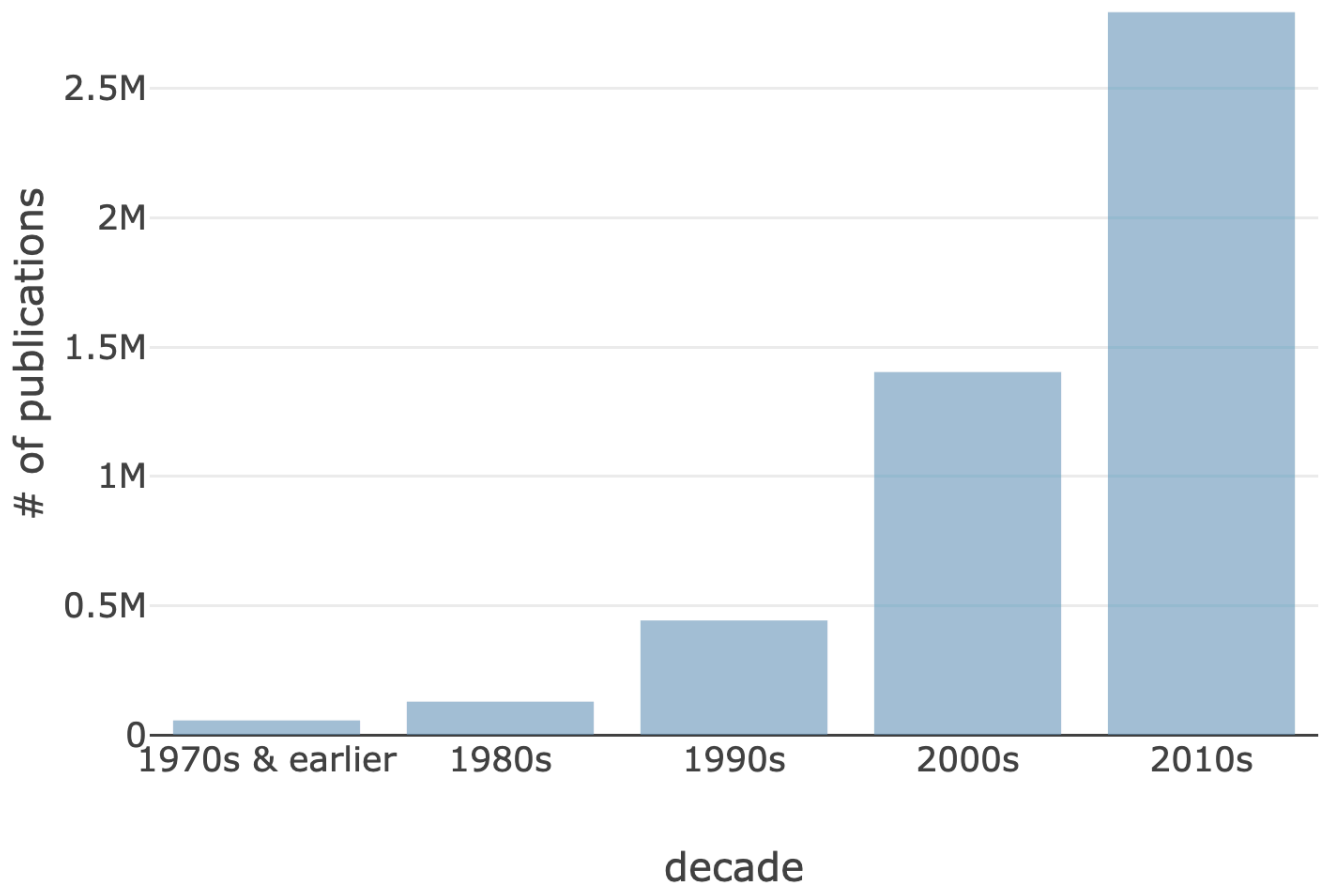
publications by conferences



As can be observed by this histogram, the great majority of conferences have published between 1 and 199 publications. This is followed by conferences that have published between 200 and 599 publications and, subsequently, by conferences that have published between 600 and 1199 publications. Similar to the trend observed in journals, we can see that few conferences have published a large number of publications.

The *co-authors.html*, *conferences.html*, *journals.html* and *years.html* files consist of graphical representations of the bins in histogram format.

publications by years



As can be observed in this histogram, the great majority of publications have been published in the current decade (approaching 2.8 million). This is followed by the previous decade (~1.4 million). The clear trend is that the number of publications is more than doubling through each decade.