Data in brief 25 (2019) 104382

Contents lists available at ScienceDirect

Data in brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A retrospective overview of International Collegiate programming contest data

Rick H. de Boer, Cassio P. de Campos*

Utrecht University, the Netherlands

ARTICLE INFO

Article history:
Received 27 May 2019
Received in revised form 29 July 2019
Accepted 2 August 2019
Available online 12 August 2019

Keywords:
Programming
Programming competition results
International Collegiate programming contest
Competition data

ABSTRACT

The International Collegiate Programming Contest[1] is an annual, multi-tier competition held amongst college students on a global scale, with world championships every year. Last year alone, around fifty thousand students from three thousand universities participated in ICPC regional competitions. Because of its significant size involving a lot of talent and skillful people, multiple stakeholders are interested in the competition. Each of the competitions results in scoreboards, containing valuable data about the performance of teams. This data however is, up till now, never collected and stored in an open and free repository. The ICPC does keep track of the basic information such as teams' names and their final scores, but more detailed information has remained scattered across the internet. This paper describes the data collected and cleaned from the European, Latin-American, North American, South Pacific and World Finals from 2012 to 2018, opening up research opportunities for an in-depth look into the programming competitions.

1. Data

The data consists of a collection of ICPC programming competition results. It contains information about teams and their scores for the problems that were posed to them and which they (tried to)

* Corresponding author.
E-mail address: cassiopc@acm.org (C.P. de Campos).
[1] https://icpc.baylor.edu.

https://doi.org/10.1016/j.dib.2019.104382

Specifications Table

Subject	Computer science – Computer Science (general)
Specific subject area	Applied Computer Science, more specifically, the context of Programming competitions. The competition for which data is gathered is the International Collegiate Programming Contest results.
Type of data	Competition results in comma separated files.
How data was acquired	Internet search, expert knowledge and extensive cleaning.
Data format	Structured: raw data from several sources has been formatted in the same structure and missing information was completed as needed Partly analyzed: part of data is used for a master thesis
Parameters for data collection	The consideration for which competition to include in the dataset was made based on three factors; the online availability of the data, the completeness and the same ruleset as specified by the global ICPC foundation being used within the competition. Only the European, Latin-American, South-Pacific, World Finals and some of the North American competitions adhered to these requirements.
Description of data collection	Data was collected from publicly available sources in the form of scoreboard tables. Sources for each table, representing a single year from a single competition, are all listed in the Appendix. The high level ETL (Extraction, Transform and Load) process of the data is shown in Fig. 4 and described in more detail in the Experimental Design section.
Data source location	List of URL's for the sources is included in the Appendix.
Data accessibility	The data is permanently stored at Mendeley Data under the https://doi.org/10.17632/5k7xtf582g.1, Other material such as visualizations for each competition year and scripts used to process the obtained data are available at https://github.com/RickdeBoer/ICPC-Scoreboards/

Value of the data

- Provides opportunities to get a retrospective of past competitions and opens possibilities for research on competition results.
- Competitions have become a major form of evaluation of research quality, acquired knowledge and skills. Performing well in these programming competitions require all the most valuable skills that are sought by major technological companies. Therefore, the data can be used to help in understanding whether universities are providing means for students to obtain such skills, benefiting all parties involved.
- With the inclusion of multiple competition and years, and also several attributes of each team, insights can be gained from comparing and analyzing over multiple dimensions. This leads to better understanding of the competition and could encourage the further improvement of it internally.
- The data is an extension of the publicly available ICPC data[2] (therefore it can be combined relatively easy), thereby directly providing more information than has been available until now.

solve. This dataset covers the ICPC regions of Europe, North America, Latin-America, South Pacific and the World Finals, from 2012 up till 2018. Around fifty thousand students from three thousand universities participated in the ICPC events in 2018 [1]. The overall data structure is shown in Fig. 1.

As shown in Fig. 1, the data is divided by topic for easy extraction and/or combination of desired information. In this context, an *Entry* represents a team and all its information, which has entered in a (single) competition. This also includes the team's final rank for that competition, their final score, consisting of the number of problems solved, and total time taken. This is the time elapsed from the beginning of a contest till the first accepted submission of a problem, accumulated for each problem, including a penalty for every additional attempt. Note that, in contrast to the team information normally present in the public ICPC data, team names are included in this dataset but cannot be assumed to be completely identical, as the data of the original sources were often not the same as the official ICPC data. This has no direct limitations, as a team's name is not an essential information as for any subsequent analysis, since it is not directly associated with any other information (even the same name can be used by different teams and/or different team members). Exploring the data structure in the

[2] https://icpc.baylor.edu/regionals/results.

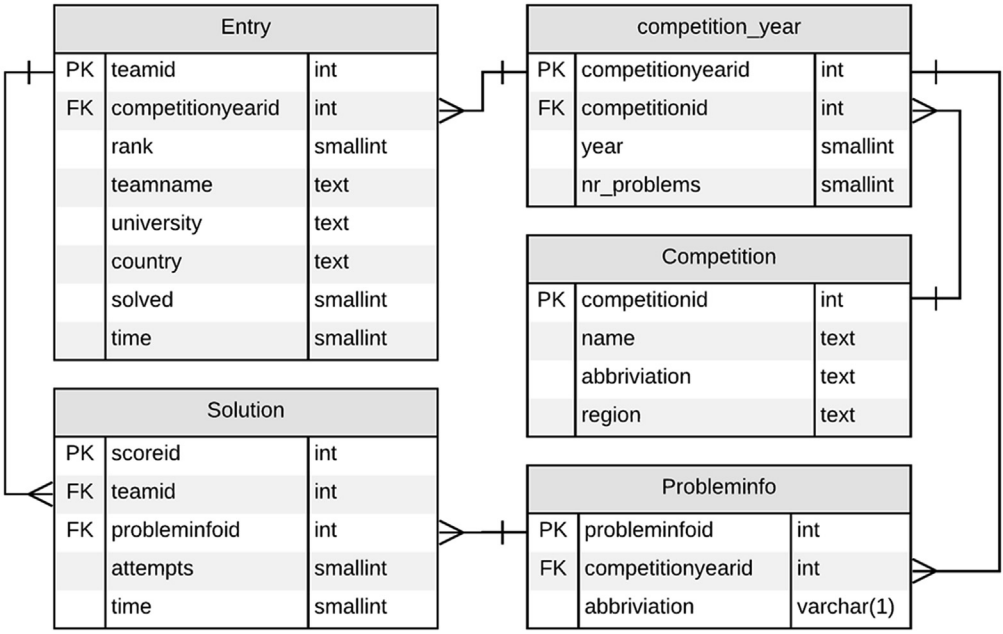

Fig. 1. Entity Relationship Diagram for the structure of the data.

figure further, a *competition* stores some meta-information, such as its region, the years it was held (i.e. the years that are present in the data) and the size of the problem set for each year. Finally, a *solution* represents all input from a single team for a single problem, where the 'attempts' are the number of times a team tried to solve a problem and the time is the total time it took to solve. This means that, if no time and only a number of attempts is present, a team did not solve that problem, and if no entry exists for a combination of a team and problem, that particular team has made no attempts on that problem. All five tables in this diagram are separate files in the dataset, which can be combined using the corresponding identifiers.

In total, the data consists of 15141 team rows which provided 60544 solution rows. These teams participated in 129 unique matches, from 23 different competitions (aggregated into 19 in this dataset) of 5 distinct regions; Europe, Latin-America, North America, South Pacific and World Finals. Those matches had 1362 problems in total, which are 10.558 problems on average. Note that not all competitions from the North American region are part of this dataset, because some of them were not publicly available. Also, the years 2012−2014 for the Mid-Atlantic USA Regional Contest and 2014 from the South-East USA Regional contest are missing because they are not (publicly) available online.

To give an overview of the distribution of participants over the regions and to indicate the popularity and scale of the competition, Table 1 shows the total number of participating teams for each region. This information is at regional level; some regions may have many more participants in prior/qualifying levels, but these are not present in this dataset: only regional finals are considered. There can be seen there that on average, the World Finals is the biggest region in terms of absolute number of teams entered, followed by Europe and North America. Table 2 further details the participant distribution, showing the most frequent countries (out of the 89 total) where the teams originate from, and all other participating countries are illustrated in Fig. 2.

The team's country of origin is often directly related to competition region and its university, as teams normally compete in their local competitions. Large countries are represented the most in this dataset. Besides the large presence of Russian teams, the top 10 is mostly filled with countries coming from Latin America, such as Mexico, Chile and Brazil. This data is essentially a more detailed

Table 1
Total number of teams and competitions per region present in the dataset.

Region	Number of teams	Number of competitions	Average number of teams per competition
Europe	3950	5	112
Latin America	2735	1[3]	78
North America	7345	11	95
South Pacific	233	1	33
World Finals	877	1	125

Table 2
Top 20 most frequently represented countries in this dataset.

Country	Nr. of Teams	Country	Nr. of Teams
USA	7024	ARG	239
RUS	1013	KAZ	226
CAN	485	POL	203
MEX	484	AUS	194
BRA	470	CUB	192
BOL	300	DEU	192
UKR	297	GBR	190
COL	295	FRA	168
PER	261	ROU	161
CHI	241	NLD	147

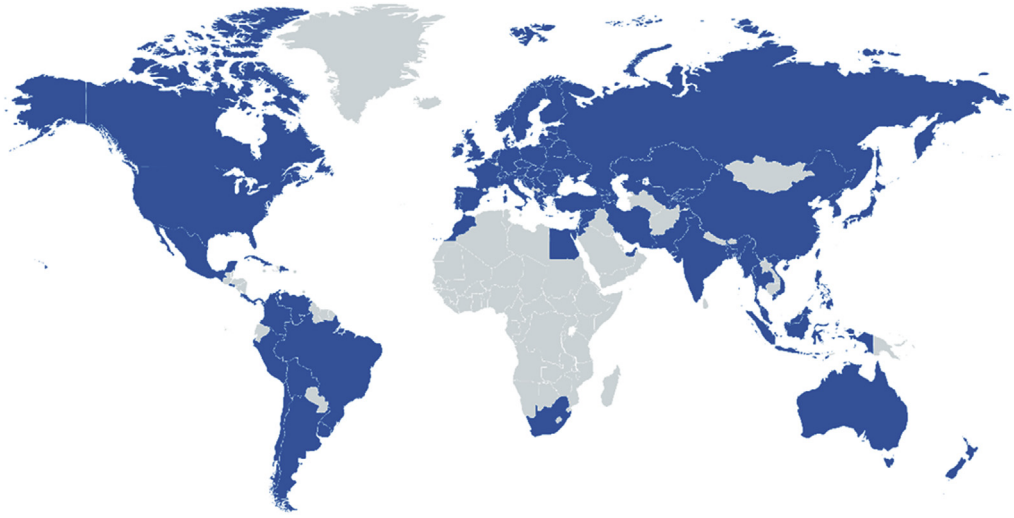

Fig. 2. Country of origin from teams' universities present in this dataset.

version of Table 1 as country and region are perfectly correlated with a Cramer's V of 1 if you exclude the World Finals; all regions have unique countries, meaning that teams only participate in local competitions.

[3] All five regional competitions of the Latin American region are aggregated, as these competitions are held at the same time and use the same set of problems. This essentially makes it a single big competition with multiple sites; hence it is being viewed as one in this context.

Besides information about the background of teams, the main part of the data are the scores of teams that attempted to solve problems. The data is suited for comparison as well for people interested in a specific competition or year. Table 3 shows an example of some basic statistics, which can be easily calculated for all other competitions as well. Here you can see indications of some problems being more difficult than others, as they (on average, maximum and/or total) have more attempts or a higher average timepoint of solving. Furthermore, Table 4 shows more statistics related to the time in which problems were solved, as another example of the information that can be learned from the data. Finally, Fig. 3 shows a way to visualize the solutions in a particular competition and year, where you can see how often and at which timepoints each problem was solved. Visualizations for all years and competitions, and code for producing these visualizations are available through the corresponding links.

2. Experimental design, materials and methods

As is illustrated in Fig. 4, the process of data acquisition starts with locating the scoreboard sources. This is a significant task in itself, as each competition has its own website and all the information is often not structured the same way. For older and missing data, web archives were consulted. Once the correct webpage has been found, the data is downloaded in the format it is provided in, either a HTML or PDF page. These pages are then subsequently processed into CSV files, either manually or with the help of tools. The transforming of the scoreboard files was the next significant step. Cleaning was done by automatically going over all rows one by one with and carrying out several actions, based on the given data and its structure. Note here that, although often almost all steps needed to be carried out, not every step was necessary for each single data set entry (it depended on the quality and format of the available data). For instance, molding the columns into the

Table 3
Example statistics from the Northeast European Regional Contest 2017.

Problem	Max. attempts	Avg. attempts	Total attempts	Min. time	Max. time	Avg. time
A	44	4.59	680	42	297	199.13
B	29	2.70	693	17	297	95.42
C	32	4.23	884	21	298	126.67
D	25	3.23	436	66	295	196.03
E	16	2.95	786	6	284	34.97
F	15	3.16	136	70	299	216.00
G	11	3.50	28	216	290	247.67
H	1	1.00	1	N/A	N/A	N/A
I	20	3.74	161	103	296	203.80
J	6	2.36	33	115	287	199.33
K	14	4.86	68	171	289	241.67
L	12	3.60	151	94	291	201.70

Table 4
Example statistics from the Mid-Central USA Regional Contest 2018.

Problem	Max attempts to solve	Avg attempts to solve	Total attempts to solve	Solved by % of teams
A	7	226	70	26%
B	6	158	178	94%
C	10	360	18	4%
D	4	124	147	99%
E	6	183	11	5%
F	8	126	137	91%
G	3	117	21	15%
H	4	121	128	88%
I	1	100	1	1%
J	7	400	8	2%
K	10	288	23	7%

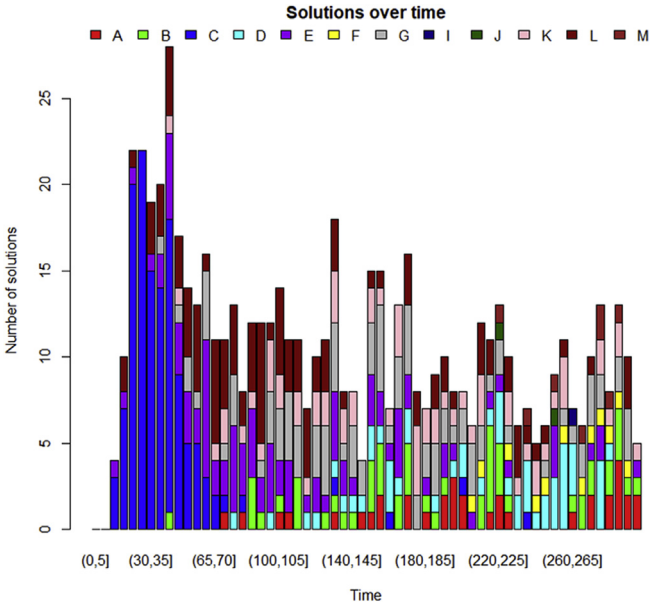

Fig. 3. Solutions over time for all problems from World Finals 2016.

same order was not always necessary, and the 'clean team information' step can consist out of multiple sub steps (such as adding, removing and restructuring data), but not all those steps were always carried out. For example, sometimes the university was not given explicitly, but only an image of the logo. Another step worth highlighting here was restructuring the scores; at least seventeen different ways of writing the scores have been found. All of those are restructured to the same format, where time penalties are excluded from the individual solution entries (so the time represents the pure time taken in that table).

After a cleaned file for a certain year and competition was ready, the now roughly cleaned files could be transformed and loaded into the database, marking the start of phase two. The missing information about teams was filled by using external sources, which mostly consisted of adding the country in which the university of the teams is situated. The final and most time-consuming task was to correct and (cross-)validate all information. These iterative steps were carried out several times, because oftentimes, new incorrections or impurities were found. The great majority of information about universities was cleaned extensively, very often manually, because university names were non-trivially

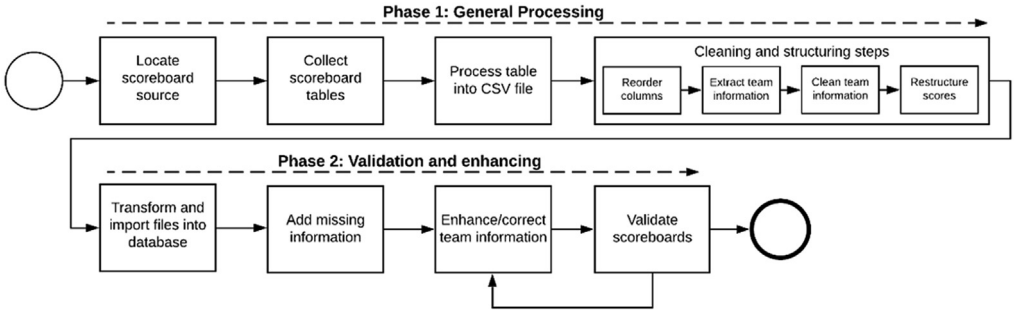

Fig. 4. Data processing steps applied to each competition.

abbreviated in the data and not always written in the same way (e.g. 'University of Utrecht' and 'Utrecht University' were to different names for the same university). The goal here was to have the names of universities written in a uniform way. A note here is that, as the original sources were used, these names are also often in the language of the source, e.g. a Brazilian university such as the 'University of Brasilia' is often written in Portuguese as 'Universidade de Brasília' (following the convention of the official ICPC results). This also means that the number of different universities in this dataset is not necessarily exactly the same as the number of distinct universities present in there, as it is possible that the same university has been written in multiple languages for different competitions. This iterative way of enhancing and validating was repeated until the data was found to be correct and complete.

3. Limitations

This dataset is limited in two ways. First, the precise recreation of this dataset is difficult and time consuming, as some processing steps have been done semi-manually or even completely manually, because not all data processing could be automatized. Second, as not all results for some regions and competitions were publicly available, only some competitions are included in this set. This could cause issues with the generality of the competition results as a whole, which must be kept in mind when interpreting the data.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Data sources

In this section, the original sources from all gathered data are detailed. This is given to ensure some reproducibility and traceability of this research. However, this list is not continuously updated, and links could therefore become outdated. The list is confirmed to be working on the 30th of April 2019.

Table A.1
World final data sources

Competition	Year	Source
World Finals	2018	https://web.archive.org/web/20180424212750/https://icpc.baylor.edu/scoreboard/
	2017	http://static.kattis.com/icpc/wf2017/
	2016	http://board.acmicpc.info/icpc2016/board.php
	2015	http://board.acmicpc.info/icpc2015/board.php
	2014	http://board.acmicpc.info/icpc2014/board.php
	2013	http://board.acmicpc.info/icpc2013/board.php
	2012	http://board.acmicpc.info/icpc2012/board.php

Table A.2
European regional data sources

Competition	Year	Source
Central Europe	2018	https://contest.felk.cvut.cz/18cerc/rank.html
	2017	https://contest.felk.cvut.cz/17cerc/rank.html
	2016	http://cerc.hsin.hr/2016/
	2015	http://cerc.hsin.hr/2015/
	2014	https://cerc.tcs.uj.edu.pl/2014/ranking.html
	2013	https://cerc.tcs.uj.edu.pl/2013/ranking.html
	2012	https://cerc.tcs.uj.edu.pl/2012/ranking.html
Northeastern Europe	2018	https://neerc.ifmo.ru/archive/2018.html
	2017	https://neerc.ifmo.ru/archive/2017.html
	2016	https://neerc.ifmo.ru/archive/2016.html
	2015	https://neerc.ifmo.ru/archive/2015.html
	2014	https://neerc.ifmo.ru/archive/2014.html
	2013	https://neerc.ifmo.ru/archive/2013.html
	2012	https://neerc.ifmo.ru/archive/2012.html
Northwestern Europe	2018	http://2018.nwerc.eu/
	2017	http://2017.nwerc.eu/
	2016	http://2016.nwerc.eu/
	2015	http://2015.nwerc.eu/
	2014	http://2014.nwerc.eu/
	2013	https://2013.nwerc.eu/en/results/scoreboard/
	2012	https://2012.nwerc.eu/en/results/scoreboard/
Southeastern Europe	2018	http://acm.ro/
	2017	https://web.archive.org/web/20171025160647/http://acm.ro
	2016	http://acm.ro/2016/
	2015	http://acm.ro/2015/
	2014	http://acm.ro/2014/
	2013	http://acm.ro/2013/
	2012	http://acm.ro/2012/
Southwestern Europe	2018	https://swerc.eu/2018/theme/scoreboard/public/
	2017	https://swerc.eu/2017/theme/results/official/public/
	2016	https://swerc.eu/2017/theme/cached/2016/
	2015	https://swerc.eu/2017/theme/cached/2015/
	2014	https://swerc.eu/2017/theme/cached/2014/
	2013	https://swerc.eu/2017/theme/cached/2013/
	2012	https://swerc.eu/2017/theme/cached/2012/

Table A.3
Latin American regional data sources (Note that all sub-competitions are stored in the same scoreboard at the same website)

Competition	Year	Source
Latin America	2018	http://maratona.ime.usp.br/resultados18/
	2017	http://www.bombonera.org/oldboards/score2017/score/
	2016	http://bombonera.org/oldboards/score2016/
	2015	http://bombonera.org/oldboards/score2015/
	2014	http://bombonera.org/oldboards/score2014/
	2013	http://bombonera.org/oldboards/score2013/
	2012	http://bombonera.org/oldboards/score2012/score2012/

Table A.4
South Pacific regional data sources

Competition	Year	Source
South Pacific	2018	http://public.webdev.aut.ac.nz/ACM/Scoreboards/2018/Regional/FinalScoreboard2018.html
	2017	http://public.webdev.aut.ac.nz/ACM/Scoreboards/2017/Regional/FinalScoreboard.htm
	2016	http://public.webdev.aut.ac.nz/ACM/Scoreboards/2016/Regional/FinalScoreboard.htm
	2015	http://public.webdev.aut.ac.nz/ACM/Scoreboards/2015/Regional/Scoreboard.htm
	2014	http://public.webdev.aut.ac.nz/ACM/Scoreboards/2014/Regional/Scoreboard.html
	2013	http://public.webdev.aut.ac.nz/ACM/Scoreboards/2013/Scoreboard.html
	2012	http://public.webdev.aut.ac.nz/ACM/Scoreboards/2012/Scoreboard.html

Table A.5
North American regional data sources (Note that not all competitions, but only the ones available online, are listed here)

Competition	Year	Source
East-Central NA	2018	https://ecna18.kattis.com/standings/standalone
	2017	https://ecna17.kattis.com/standings/standalone
	2016	https://ecna16.kattis.com/standings/standalone
	2015	https://ecna15.kattis.com/standings/standalone
	2014	https://web.archive.org/web/20160828011045/http://acm-ecna.ysu.edu:80/PastResults/2014/standings.html
	2013	https://web.archive.org/web/20131115071843/http://icpc01.cc.ysu.edu:80/scoreboard/
	2012	https://web.archive.org/web/20150822192024/http://acm.ashland.edu:80/2012/standings.html
Greater NY	2018	http://acmgnyr.org/year2018/standings.shtml
	2017	http://acmgnyr.org/year2017/standings.shtml
	2016	http://acmgnyr.org/year2016/standings.shtml
	2015	http://acmgnyr.org/year2015/standings.shtml
	2014	http://acmgnyr.org/year2014/standings.shtml
	2013	http://acmgnyr.org/year2013/standings.shtml
	2012	http://acmgnyr.org/year2012/standings.shtml
Mid-Atlantic USA	2018	https://mausa18.kattis.com/standings
	2017	https://mausa17.kattis.com/standings
	2016	https://web.archive.org/web/20161109015551/http://midatl.radford.edu:80/scoreboard/summary.html
	2015	https://web.archive.org/web/20160118135544/http://midatl.radford.edu:80/scoreboard/summary.html
	2014	https://web.archive.org/web/20150519074453/https://www.cs.odu.edu/~zeil/icpc/scoreboard2014.html
	2013	https://web.archive.org/web/20140401093853/http://www.radford.edu:80/~acm/midatl/2013_scoreboard.html
	2012	http://midatl.radford.edu/docs/scoreboard/summary.html
Mid-Central USA	2018	https://mcpc18.kattis.com/standings
	2017	https://mcpc17.kattis.com/standings
	2016	https://mcpc16.kattis.com/standings
	2015	https://mcpc15.kattis.com/standings
	2014	N/A
	2013	N/A
	2012	N/A
NA Invitational	2018	https://naipc18.kattis.com/standings/standalone
	2017	https://naipc17.kattis.com/standings/standalone
	2016	https://naipc16.kattis.com/standings/standalone
	2015	https://naipc15.kattis.com/standings/standalone
	2014	http://naipc.uchicago.edu/2014/scoreboard-final-onsite.html
	2013	http://icpc.cs.uchicago.edu/invitational2013/board_final.html
	2012	http://icpc.cs.uchicago.edu/invitational2012/scoreboard.html
North-Central NA	2018	https://ncna18.kattis.com/standings
	2017	https://ncna17.kattis.com/standings
	2016	http://cse.unl.edu/~upe/contest/
	2015	http://cse.unl.edu/~upe/contest/
	2014	http://cse.unl.edu/~upe/contest/
	2013	http://cse.unl.edu/~upe/contest/
	2012	http://cse.unl.edu/~upe/contest/
Pacific North-West	2018	http://acmicpc-pacnw.org/scoreboard/2018/index1.html
	2017	http://acmicpc-pacnw.org/ProblemSet/2017/index1.html
	2016	https://web.archive.org/web/20170111143951/http://www.acmicpc-pacnw.org:80/scoreboard/index1.html
	2015	http://acmicpc-pacnw.org/ProblemSet/2015/index1.html
	2014	https://web.archive.org/web/20141224050227/http://www.acmicpc-pacnw.org:80/scoreboard/index1.html
	2013	http://acmicpc-pacnw.org/ProblemSet/2013/index.html
	2012	http://acmicpc-pacnw.org/Standings/2012/index.html
Rocky mountain	2018	https://rmc18.kattis.com/standings
	2017	https://rmc17.kattis.com/standings

(*continued on next page*)

Table A.5 (*continued*)

Competition	Year	Source
	2016	https://rmc16.kattis.com/standings
	2015	https://rmc15.kattis.com/standings
	2014	https://org.coloradomesa.edu/~wmacevoy/rmrc/2014/scoreboard.html
	2013	https://org.coloradomesa.edu/~wmacevoy/rmrc/2013/scoreboard_byrank.html
	2012	https://web.archive.org/web/20130913013048/http://org.coloradomesa.edu:80/acm/rmrc/2012/scoreboard_byrank.html
South-Central USA	2018	http://ld2018.scusa.lsu.edu/standings-contest/
	2017	http://ld2017.scusa.lsu.edu/scoreboard-regional/
	2016	http://ld2016.scusa.lsu.edu/scoreboard-regional/
	2015	http://ld2015.scusa.lsu.edu/scoreboard-regional/
	2014	http://ld2014.scusa.lsu.edu/scoreboard-regional/
	2013	http://ld2013.scusa.lsu.edu/scoreboard-regional/
	2012	http://ld2012.scusa.lsu.edu/scoreboard-regional/
South-East USA	2018	https://ser.cs.fit.edu/ser2018/ser2018-results-div1.pdf
	2017	https://ser.cs.fit.edu/ser2017/ser2017-results-div1.pdf
	2016	https://ser.cs.fit.edu/ser2016/ser2016-results-div1.pdf
	2015	https://ser.cs.fit.edu/ser2015/ser2015-results-div1.pdf
	2014	N/A
	2013	https://ser.cs.fit.edu/ser2013/ser2013_final_standingsl.pdf
	2012	https://ser.cs.fit.edu/ser2012/ser2012_scoreboard.pdf
Southern California	2018	http://socalcontest.org/history/2018/Scoreboard-2018.shtml
	2017	http://socalcontest.org/history/2017/details-2017.shtml
	2016	http://socalcontest.org/history/2016/details-2016.shtml
	2015	http://socalcontest.org/history/2015/details-2015.shtml
	2014	http://socalcontest.org/history/2014/details-2014.shtml
	2013	http://socalcontest.org/history/2013/details-2013.shtml
	2012	http://socalcontest.org/history/2012/details-2012.shtml

References

[1] The ICPC Foundation, ICPC Fact Sheet [Fact Sheet], 2018. Retrieved from, https://icpc.baylor.edu/worldfinals/pdf/Factsheet.pdf.