Noname manuscript No.
(will be inserted by the editor)

Genetic programming in the 21st century: a bibliometric and content-based analysis from both sides of the fence

Andrea De Lorenzo · Alberto Bartoli · Mauro Castelli · Eric Medvet · Bing Xue

Received: date / Accepted: date

Abstract In this work we present an extensive *bibliometric* and *content-based* analysis of the scientific literature about genetic programming in the 21st century. Our work has two key peculiarities. First, we revealed the *topics* emerging from the literature based on an *unsupervised* analysis of the *textual content* of titles and abstracts. Second, we executed all of our analyses twice. On the papers published on the venues that are typical of the evolutionary computation research community and those published on all the other venues. This view from "both sides of the fence" allows us to gain broader and deeper insights into the actual contributions of our community.

Keywords Genetic Programming, Bibliometrics, Topic modeling, Literature review, Publication habits

1 Introduction

Genetic programming (GP) as a research field has experienced significant growth since the journal Genetic Programming and Evolvable Machines (GPEM) commenced publication in year 1998. As of October 2018, the query "genetic programming" in Google Scholar returned 6890 items for the time frame 1988–1998, 24 700 for the time frame 1998–2008 and 35 400 for 2008–2018. While these rough figures are not a measure of the overall impact on the scientific community, the relevance and maturity of the GP framework cannot be questioned.

Andrea De Lorenzo, Alberto Bartoli, Eric Medvet
Department of Engineering and Architecture, University of Trieste, Trieste, Italy
E-mail: andrea.delorenzo@units.it, E-mail: bartoli.alberto@units.it, E-mail: emedvet@units.it

Mauro Castelli
NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312, Lisboa, Portugal
E-mail: mcastelli@novaims.unl.pt

Bing Xue
School of Engineering and Computer Science, Victoria University of Wellington, Wellington, New Zeland
E-mail: Bing.Xue@ecs.vuw.ac.nz

In this work, we seek to obtain an understanding of where the field has been and where it is headed, based on several forms of *bibliometric* and *content-based* analysis executed on a large collection of scientific papers published from the year 2000 onward[1]. Specifically, we considered title, abstract, publication venue, and other metadata of all the papers indexed in the Scopus database returned by the "genetic programming" search query. Such a dataset arguably provides near complete coverage of the academic contributions of our field in this century.

There are two key elements in our study. First, we revealed the *topics* emerging from the *content* of titles and abstracts based on *probabilistic topic modeling*, a text mining technique that has proven its power and effectiveness in a broad range of domains [5]. The fact that this technique is *unsupervised* provides the opportunity to glean insights about the actual content of the scientific production while minimizing the set of a priori assumptions about such content. Second, we executed *two* full sets of analyses: one on the papers published on the venues that are typical of the evolutionary computation (EC) research community, another on those published in all the other venues where the term GP appears among the topics of interest. This dual study from "both sides of the fence" allows us to evaluate the actual contributions of our research field from a broader perspective, complementary to that of our community and capable of providing important insights into the overall development and impact of the GP framework. We are not aware of any other study similar to ours.

We begin our analysis by focusing on the *number of published papers*, their distribution across publication venues, and the temporal evolution of these quantities. We illustrate the most important *topics* that emerged from the corpus, the most important keywords associated with those topics, how the interest in the topics varied over time, and the most representative documents for each topic. We then analyze the number of *citations*, its relation with the venue of publication, and its temporal evolution. Finally, we carry out a *geographical* analysis based on the papers' authors' countries of affiliation, as well as the distribution of topics across countries.

The main findings of our analysis can be summarized as follows:

1. Papers on GP are more frequently published at venues that are outside of the traditional circles of the EC community.
2. The impact, assessed in terms of the number of citations, of papers published in EC venues is roughly the same as those published in non-EC venues.
3. The interest in different research topics related to GP varies over time.
4. Though the countries that initiated the field (US, UK) are still offering important contributions, new countries are emerging as fertile ground for GP research.

2 Related work

There have been a large number of publications on GP, including books, journal papers, conference proceeding papers, and online resources. The GP bibliography [1] contains over 12 000 references, and using "genetic programming" as the

[1] Formally, the 21st century started on January 1, 2001, rather than on January 1, 2000, which is the starting date of the so-called "2000s century." We chose, however, to use "21st century" because we think it is a more accessible locution.

keyword to search in Google scholar yielded over 167 000 results (assessed on October 17, 2018). Considering the scope and the length of this paper, it is not possible to cover all existing literature on GP. In the next two sections, we will instead survey the most significant past reviews of GP literature (in Section 2.1) and other existing studies and methods that sought to quantitatively analyze the scientific literature (in Section 2.2).

2.1 Review of GP literature published in the past

In 2010, Langdon and Gustafson [16] surveyed the first ten years of reviews in the GPEM journal, where almost every issue includes at least one review, mainly on books. Those books covered both the theory and applications of GP. Almost all of the key books and resources on GP by then have been reviewed in GPEM, and the survey paper provided a comprehensive summary of those reviews.

An early survey paper on GP was written by Koza in 1995 [15]; it presented an introduction on genetic algorithms (GAs) and GP, and listed books and available software. The cited survey discussed various types of representations (i.e., the most commonly used standard tree-based GP, strongly typed tree-based GP, grammar-based GP, stack-based GP, linear GP, and Cartesian GP). A more recent survey on various representation of linear GP was published in 2009 [23], which discussed the strengths and weaknesses as well as the online resources available for each representation. McKay et al. [18] provided a survey on grammar-guided GP (G3P) in which they reviewed tree-based G3P, linearized G3P, and other representations that extend the context-sensitive grammar or incorporate semantic knowledge into the grammar representation. The benefits of grammar in GP were also discussed, followed by the review of the major application areas of G3P. In recent years, semantic GP has become very popular. Vanneschi at al. [29] discussed different methods in semantic GP, including methods that maintain the semantic diversity in the population, the methods that indirectly promote a semantic behavior via the survival criteria, and the semantic methods that use precise genetic operators directly on the semantics of individuals.

Classification and symbolic regression are two main applications of GP [17]. There is no comprehensive survey on GP for symbolic regression, probably due to the fact that the number of papers on this topic is too big to have a survey to review all the important work. Espejo et al. [11] surveyed existing work on GP for classification, which covered typical work on using GP for selecting relevant features, constructing high-level features, extracting classifiers, and learning ensembles. It showed that GP has achieved promising results on those aspects on a wide range of real-world problems. The ability to simultaneously evolve a classifier and perform feature selection is one of the salient characteristics of GP, and the survey paper by Xue and coauthors [33] discussed existing works in this area. Another survey [27] has shown that GP is the most powerful approach for feature construction and reviewed various GP methods for this task.

In addition, GP uses a flexible and powerful representation of candidate solutions, which naturally makes it a hyper-heuristics approach. A survey paper on using GP to evolve production scheduling rules in planning and scheduling was published in 2016 [8]. Furthermore, a survey on genetic improvement (i.e., automatically searching for improved versions of existing software) has been presented in 2018 [24], which showed that 96% of the core publications in the area used GP for genetic improvement.

Besides the promising performance of GP, there are also some limitations and issues, some of which have been discussed by Dabhi and Chaudhary [9]. Specifically, this survey paper discussed the computational efficiency, premature convergence, the model-building ability, the generalization issue, and constant creation in GP. Meanwhile, this paper also reviewed different methods for tackling the above issues, along with discussions about their advantages and disadvantages.

A position paper [17] entitled "Genetic Programming Needs Better Benchmarks" that was presented at the Genetic and Evolutionary Computation Conference (GECCO) in 2012 argued that there are many problems with the existing benchmarks that were widely used in the GP community, and proposed to have a standardized benchmark suite. Following [17], a community survey was conducted [32]. The survey results showed that the community in general supported the creation of new benchmark suites. A number of alternative benchmarks were also recommended for major application areas of GP, including symbolic regression problems, Boolean problems, classification problems, planning and control problems, and algorithmic programming problems.

2.2 Quantitative methods for literature analysis

Quantitative analysis on literature is an effective way to analyze research activities and discover trends and hidden relationships among existing work. Such analysis can provide support for the development of future research. Here we briefly review typical quantitative literature analysis articles.

A network-based analysis was conducted in [14] to investigate the evolution of different scientific fields; in this effort, scientific concepts were presented by nodes in the network and two nodes were linked if they have been referenced in the same publication. The analysis on data from 1985 to 2006 discovered the evolution patterns of different fields and identified communities that can map to known research areas.

Methods for quantitative analysis of the textual content of the literature are increasingly based on *unsupervised* approaches, as is usually done for examining and processing large collections of web pages [19,21,28]. Blei [5] discussed probabilistic topic modeling, a text mining technique that can be used for discovering and annotating documents in large archives. As a running example, he presented the results of the analysis of issues of the journal Science spanning more than one century using Latent Dirichlet Allocation, a form of probabilistic topic modeling. We used the same technique in the present study.

Social network analysis (SNA), which seeks to discover relationships between authors in particular, is a common method to conduct quantitative literature analysis, but SNA does not take into account the actual content of the papers. De Nart et al. [10] developed an approach combining SNA and content-based analysis. The method used an automatic knowledge extraction tool to finely model the research topics and was tested to be effective on 7000 papers in computer science and ICT.

In 2018, Bao et al. [3] performed a comprehensive analysis of the publications related to soft robotics using data from the Science Citation Index Expanded database from 1990 to 2017. Based on the statistics of different criteria, such as keywords, citation, and h-index, this analysis showed overviews of the field in terms of the main contributions, cooperation patterns, the most productive journals, and others.

In the EC area, journals and conferences can provide some high-level overall statistics on their publications. In particular, the report from the GECCO 2017 conference [2] showed the statistics and collaboration network between authors based on the data from 2005 and 2017. The reports also focused on the evolution of the GP track in GECCO over the course of 13 years: it highlighted that the track became slightly smaller, probably because there are more tracks in the conference, which also cover GP-related work.

In the already-cited survey paper [16] on the ten years of reviews in GPEM, Langdon and Gustafson also showed different statistics about GP based on the data from the GP bibliography [1]. The statistics were shown in terms of the number of entries from different types of publications, the number of people active in GP, and the distribution of the number of authors. The use of the GP bibliography over different countries, the most downloaded papers and authors, the co-authorship community, and the use of Internet and freeware in the GP community are also presented. Further, the authors pointed out that there have been many GP application papers that were published in non-computer science venues and are not included in the statistics. In our study, we extend and update this kind of analysis.

3 Data collection

We built a corpus of scholarly articles related to GP using the search engine provided by Scopus[2], an abstracts and citations database of peer-reviewed literature maintained by Elsevier. We opted for using Scopus, rather than other similar sources as, e.g., Google Scholar or Web of Science, because of (a) the availability of an API to programmatically query the search engine and (b) its good, although not perfect, coverage [13, 22]. For consistency, we will use the term *document* from now on to refer to all kinds of scholarly articles (e.g., journal articles, conference papers, book chapters).

We obtained the documents by submitting the query "genetic programming" (enclosed by double quotes) to Scopus to be searched in the title, abstract, and keywords fields, and limiting the search to documents published since the year 2000 (included). We posed no other constraints on the results, notably not on the document type. From the results of this search, we removed 344 documents whose authors' details were missing. The resulting corpus contains 10 233 documents, each with the following information: title, abstract, authors, authors' affiliations, number of citations, year of publication, publication venue[3], publication volume, and document type.

Concerning documents of the "conference paper" type, we noticed that many of them, related to a large set of conferences, had the "source title" field set to "Lecture Notes in Computer Science" instead of to the actual conference name. In order to address this issue, we applied the following procedure to each document: (i) we retrieved the volume number associated with the document; (ii) we

[2] `https://www.scopus.com`

[3] The publication venue is shown in the "source title" field of Scopus results.

Table 1 Numbers of documents for each type of venue.

Venue type	Documents	Venues
Non-EC	7387	2927
EC	2846	50

performed a search on Scopus for all documents with the same volume number whose "document type" was equal to "conference review;" (iii) we took the title of the first result as the name of the conference. This allowed us to associate the actual name of the conference with each document of type "conference paper."

We associated each document in our corpus with a label indicating whether the document was published in a publication venue that mostly focuses on EC. We built the list of EC venues in two steps: first, we considered all of the venues (journals and conferences) listed on the SPECIES website[4]: the list contains the most important publication venues related to the EC community. Then, we expanded this list by including books, book chapters, and other less-known conferences related to EC that include at least one of the following strings in the name of the venue: "evolution," "evolutionary," "genetic programming," and "genetic algorithm." We manually inspected the resulting venues in order to avoid including unrelated venues.

Table 1 reports the number of EC and non-EC venues and the corresponding number of documents.

It can be seen that the number of non-EC venues is much larger than the number of EC venues. We were not able to manually verify all the non-EC venues: we believe, however, that their large number is consistent with the fact that GP-related contributions are distributed among many different disciplines and communities.

It is worth noting that the way we collected the data may have some limitations, both in the corpus building part and in the criterion for determining the EC venues.

Concerning the former, as for any information retrieval (IR) system, there might be two kinds of error in the corpus: (a) it could include some not relevant (i.e., not GP-related) documents, and/or (b) it could fail to include some relevant documents. Those errors might be measured in terms of precision and recall, respectively, two quantitative indexes suitable for assessing IR systems. However, measuring the actual precision and recall based on manual inspection of each document is practically unfeasible, due to the scale of the problem, in particular for recall—one should, in principle, read *all* of the scholarly articles. We carefully analyzed a small, randomly chosen subset of the collected documents and, also based on our experience, we concluded that possible defects of the corpus do not affect the considerations that are later drawn in this study.

Concerning the criterion for identifying EC venues, we do acknowledge that our choice is arbitrary. In fact, it reflects the intrinsic subjectivity of this kind of choices: different scholars from different parts of the community might consider the same venue as EC or non-EC. We attempted to devise a criterion which, at the same time, builds on the experience of the most acknowledged members of

[4] `http://species-society.org`, accessed on September 2018. SPECIES is a non-profit association that "aims to promote evolutionary algorithmic thinking within Europe and wider, and more generally to promote inspiration of parallel algorithms derived from natural processes".

the EC community and can be concisely described. In these terms, the case of the "Applied Soft Computing" (ASOC) journal is significant. On one hand, the journal focus is on "research in application and convergence of the areas of Fuzzy Logic, Neural Networks, Evolutionary Computing, Rough Sets and other similar techniques." On the other hand, ASOC is not listed by SPECIES and its journal title does not include any of the above-mentioned strings. Nevertheless, our corpus contains 91 GP-related documents published in ASOC.

Finally, we decided not to use the GP bibliography [1] because for a large number of its entries it does not provide all of the information that we used for this study, most notably the abstract. Moreover, we are particularly interested in the documents tailored to or built from outside the GP community, and we argue that authors of those documents might not have known this bibliography, which could eventually have created a bias in its composition.

4 Results and discussion

This section discusses the main findings of the analysis performed over the collected corpus. The discussion takes into account different dimensions. Section 4.1 analyzes the *distribution of GP documents* across *venues*, *document types*, and its *temporal evolution*. Section 4.2 summarizes the findings of an analysis of the *content* of the corpus based on the unsupervised extraction of their *topics*. Section 4.3 takes into account the *number of citations*. Finally, Section 4.4 is devoted to a *geographical* analysis based on the documents' authors' country of affiliation.

4.1 Number of documents

4.1.1 EC venues

Table 2 reports the number of corpus documents (CD) published in the top ten EC venues (ranked according to their CD). The table also shows other information about each venue: the perceived quality in terms of 5-year impact factor (IF), for journals, or CORE rank[5], for conferences; and the volume of the venue in terms of the number of *all* papers, i.e., not only the ones included in our corpus, published yearly (YD). For the three indexes (CORE rank, IF, and YD), we show the last available value for each venue.

According to the results displayed in Table 2, GECCO, a CORE Tier A annual conference, is the main venue for GP-related documents, with 911 documents published since the year 2000. It is interesting to highlight that approximately 72% of the documents are published in three conference venues, with GECCO totaling 32% of the whole amount: these conferences (GECCO, CEC, and EuroGP) are, by far, the most important venues for disseminating research dealing with the GP.

[5] Provided by the Computing Research and Education Association of Australasia, it is one among A* (best), A, B, and C (worst), `http://www.core.edu.au/conference-portal`.

[6] The full name of this conference is "International Conference on Soft Computing: Evolutionary Computation, Genetic Programming, Fuzzy Logic, Rough Sets, Neural Networks, Fractals, Bayesian Methods."

Table 2 The top ten EC venues according to the number of corpus documents (CD column). The Type column shows the type of the venue: "J" for journal, "C" for conference. The Rank/IF column shows, for conferences, the CORE rank, if available; for journals, it shows the 5-year impact factor. The YD column shows the number of documents published in the last year by the venue.

	Venue name	Type	Rank/IF	CD	YD
1	Genetic and Evolutionary Computation Conference (GECCO)	C	A	911	193
2	IEEE Congress on Evolutionary Computation (IEEE CEC)	C	B	590	342
3	European Conference on Genetic Programming (EuroGP)	C	B	553	19
4	Genetic Programming and Evolvable Machines (GPEM)	J	1.446	165	25
5	IEEE Transactions on Evolutionary Computation (IEEE TEVC)	J	8.124	97	64
6	Applications of Evolutionary Computation (EvoApplications)	C	-	75	59
7	Parallel Problem Solving from Nature (PPSN)	C	A	67	81
8	International Conference on Computational Intelligence in Music, Sound, Art and Design (EvoMUSART)	C	-	55	20
9	International Conference on Soft Computing (MENDEL)[6]	C	-	53	29
10	Evolutionary Computation (EC)	J	2.388	41	22

The first journal that appears in Table 2 in fourth place is Genetic Programming and Evolvable Machines (IF = 1.446), with 165 documents. In fifth place it is possible to find another reputable journal in the GP field, IEEE Transactions on Evolutionary Computation (IF = 8.124) with 97 documents. Finally, the top ten is completed by another journal, Evolutionary Computation, with 41 documents. All in all, Table 2 highlights that conferences are the preferred venues for disseminating GP-related research, with GECCO standing out among all venues with its 33% of documents.

4.1.2 Non-EC venues

Table 3 reports the number of corpus documents published in the top ten non-EC venues, along with the same other information of Table 2.

The foremost finding, which is derived from the observation of Table 3, also in comparison with Table 2 (i.e., the figures for EC venues), is that when GP-related research is to be disseminated outside EC venues, researchers tend to prefer journals than conferences. In fact, six out of the ten top non-EC venues are journals. A possible interpretation of this finding is that GP researchers who need or prefer to publish in journals seek this kind of venue outside the community because the "capacity" of EC journals is low. Those non-EC journals have reasonably good impact factors, such as ASOC with an IF of 4.004 and Information Sciences with an IF of 4.832.

Another interesting observation concerns the (non) skewness of the distribution among the top ten non-EC venues. Different from EC venues, the number of documents published in the venues of Table 3 is more uniform. The ratio between

Table 3 The top ten non-EC venues according to the number of corpus documents (CD column). The Type column shows the type of the venue: "J" for journal, "C" for conference. The Rank/IF column shows, for conferences, the CORE rank, if available; for journals, it shows the 5-year impact factor. The YD column shows the number of documents published in the last year by the venue.

	Venue name	Type	Rank/IF	CD	YD
1	Studies in Computational Intelligence (SCI)	J	1.05	110	34
2	Applied Soft Computing Journal (ASOC)	J	4.004	91	702
3	Advances in Intelligent Systems and Computing (AISC)	C	C	86	38
4	Expert Systems with Applications (ESWA)	J	3.711	82	612
5	Soft Computing (SC)	J	2.367	63	599
6	Society of Instrument and Control Engineers (SICE)	C	-	56	-
7	Information Sciences (INFORM SCIENCES)	J	4.832	55	767
8	International Society for Optical Engineering (SPIE)	C	-	55	-
9	IEEE International Conference on Systems, Man and Cybernetics (IEEE SMC)	C	B	45	717
10	Neural Computing and Applications (NCA)	J	4.213	43	575

the first and the last figure is ≈ 2.5, whereas the same ratio for EC venues (see Table 2) is ≈ 22.2. Moreover, the top ten non-EC venues account for an overall 9.2% of all the documents published in that kind of venue, whereas the percentage is 72% for just the top three EC venues.

Finally, it can be seen from the name of the venues that three of them explicitly focus (also) on applications (ASOC, ESWA, and NCA). For five of them (SCI, ASOC, AISC, ESWA, SC), the name of the venue collectively refers to a set of techniques that conventionally also includes EC: most used terms are "soft" and "intelligent" computing.

4.1.3 Temporal evolution

While the previous analysis is useful to understand the venues preferred by GP scholars and practitioners, different observations may be drawn by analyzing the evolution of the field year by year. Figure 1 shows the absolute number (left) and percentage (right) of documents in EC and non-EC venues for each year since 2000. In this figure and in all of the similar figures showing the temporal evolution of some indexes, we are not plotting the values of indexes for the years 2018 and 2019 because, due to their recency, they cannot be meaningfully commented.

By observing the left plot of Figure 1, one can see how the numbers of documents in EC and non-EC venues show a similar trend in the first half of the plot. In more detail, considering the first decade of the millennium, for both EC and non-EC venues the years between 2008 and 2010 were the ones presenting the largest numbers of documents. Focusing on the years from 2011 to 2017, it is necessary to distinguish between non-EC and EC venues. In the former case, from 2011 to 2017 it is possible to notice a growing trend, with the year 2017 presenting a number of documents in EC venues comparable to the one of 2009, the absolute best year for documents in EC venues. The figures for EC venues appear to stay flat after 2011, with a number of documents in non-EC venues comparable to the

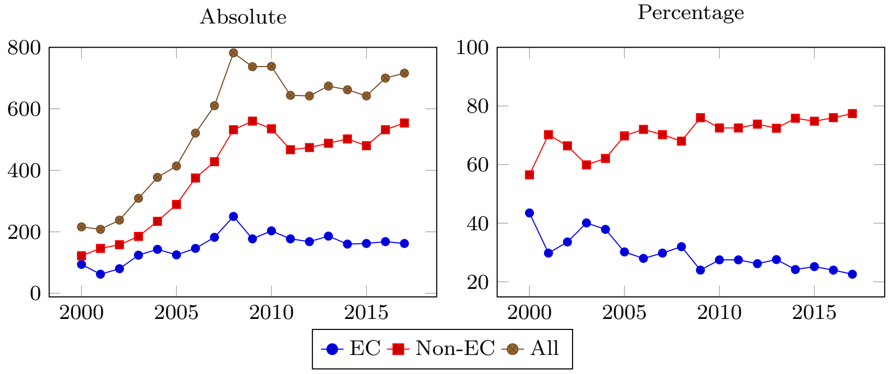


Fig. 1 Number (left) and percentage (right) of documents in EC and non-EC venues for each year.

one achieved in 2009, but significantly below the one achieved in 2008 (the best year for non-EC venues).

By looking at the right plot of Figure 1, it is interesting to note that the ratio of documents published in EC venues has slowly decreased over the time since 2000. In other words, EC community members are more and more disseminating their researches in non-EC venues. This is something particularly interesting and encouraging from the point of view of the GP community, because it shows the increasing spreading of GP as a successful tool for solving problems and as a technique from which insights applicable to different fields may be gained.

4.1.4 Document types: conference paper vs. journal article

Figure 2 plots the temporal evolution of the document type for EC and non-EC venues.

The plots, particularly the ones on the right showing values in percentage, confirm the finding of Section 4.1.2 and, in further detail, general EC venues that mainly consist of conferences. For non-EC venues, instead, the situation is evolving. It can be seen that a change occurred after 2011. In recent years, journals tend to attract more GP-related documents than conferences. Although it is not possible to ascertain to which degree this trend will be confirmed in the future, we think that this finding can be explained in terms of maturity of the GP research field.

4.2 Topic analysis

In this section we illustrate the main *topics* on which GP-related documents focused since year 2000. We use an *unsupervised* framework to minimize the bias of our analysis. Probabilistic topic modeling is a form of document clustering based on a similarity definition that attempts to capture the semantic content of a document in terms of patterns of words co-occurrences. To this end, we use *probabilistic topic modeling*, which aims at structuring the body of data in terms of naturally emerging categories rather than in terms of attributes proposed a priori by researchers [5]. Intuitively, documents that tend to contain similar patterns of words

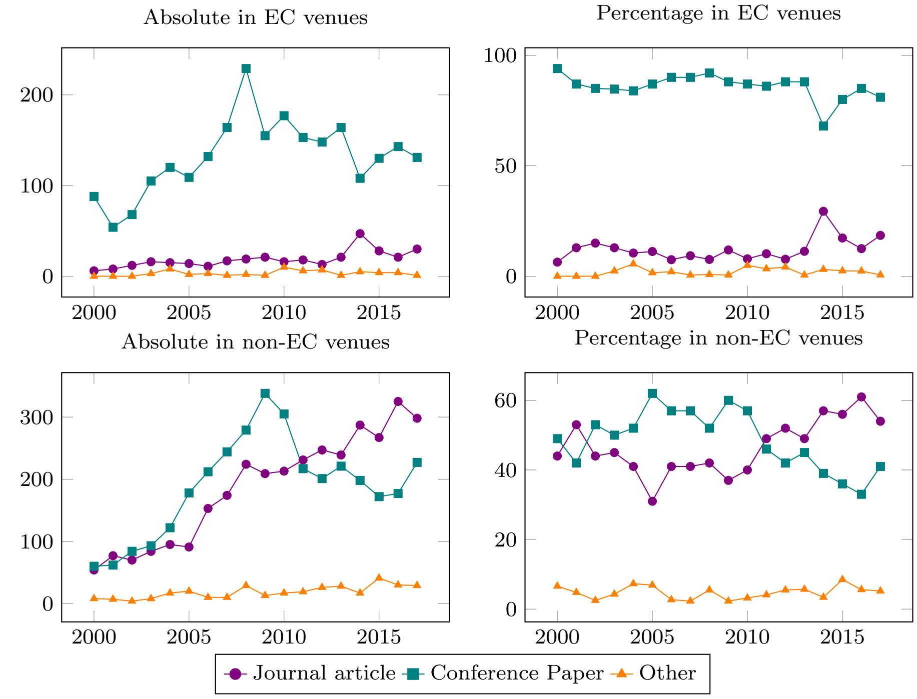


Fig. 2 Number (left) and percentage (right) of documents for the most important document types, for each year, and for EC (above) and non-EC (below) venues.

should be semantically similar in the sense that those documents should be relevant to similar topics. In order to discover and describe those patterns, the framework automatically constructs a probabilistic generative model over the available data. Being an unsupervised approach, the discovered document clusters might not be susceptible to an intuitive explanation and they might not exhibit any intuitive connection with the a priori categories expected by researchers [34]. On the other hand, a topic modeling approach to the GP scientific literature as a whole may allow uncovering insights that are impossible to obtain by manual examination.

We executed probabilistic topic modeling with the *Latent Dirichlet Allocation (LDA)* technique [7, 5]. LDA is a well-established approach for analyzing scholarly articles: for example, those published by the popular journal Science in [5] or computer science papers for venue recommendation in [20]. In this framework, the user specifies only the desired number of topics for the provided dataset. The results of an LDA execution consist of a description of the topics that emerge from the dataset and, for each document, of a description of the degree by which each document pertains to each discovered topic. A topic is described as a probability distribution over all the words, while each document is described as a probability distribution over the topics. Note that discovered topics have to be "interpreted" in the sense that they are not mapped to any predefined concept, category or word. Moreover, note that LDA does not perform a hard partition of documents over the discovered topics: each document pertains instead to all topics with different degrees.

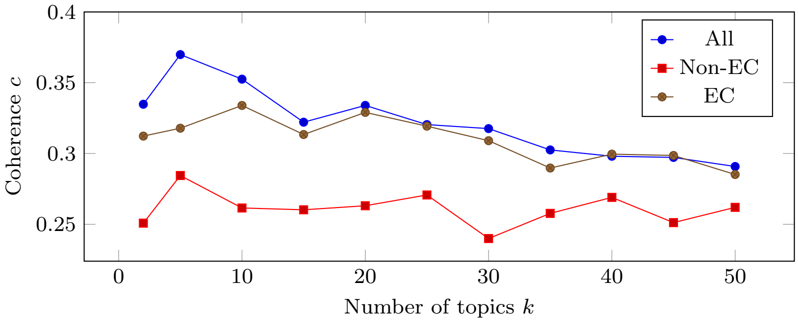


Fig. 3 Coherence c of the topics for different values of the number k of topics.

4.2.1 Choice of the number of topics and topic post-processing

Before executing LDA on our corpus we pre-processed each document by (i) concatenating its title and abstract, (ii) transforming the resulting text to lower case, (iii) removing punctuation and stop words (i.e., words that are very common and thus carry little information for natural language processing applications), and, finally, (iv) performing lemmatization (grouping together words with the same base form) and stemming (removing the end or the beginning of a word, considering a list of common prefixes and suffixes)[7]. Then, we computed the *term frequencies-inverse document frequencies* (TF-IDF) on the pre-processed corpus and removed from the documents in the pre-processed corpus all the words not included in the top 100 000 words with the largest TF-IDF. Finally, we applied the latent Dirichlet allocation (LDA) [7,5] on the resulting pre-processed corpus to discover the topics in the documents.

A key choice, and the only actual input provided by the user beyond the corpus itself, for the application of LDA is the number k of topics. A systematic and principled method for choosing value k consists in measuring the *coherence* of the resulting topics [25]: a greater coherence means a better fit of the topics over the documents. We used this method and computed the coherence c for different values of the number of topics k (ranging from 2 to 50) and for three compositions of the corpus: all of the documents, only the documents of EC venues, and only the documents of non-EC venues. Figure 3 graphically summarizes the outcome of this analysis.

It can be seen that $k = 5$ is the best choice when applying LDA to all the documents and to the corpus composed of only the documents published in the EC venues. A larger value, ($k = 10$), is instead slightly more appropriate when the corpus contains only the documents published in non-EC venues. This finding is not surprising, as these documents are likely related to different applications of GP to a broad set of diverse domains, which therefore require a larger number of topics to be described. Since we want to analyze the three corpora uniformly, however, we chose to set $k = 5$ for all of them.

[7] Both lemmatization and stemming have been done using the NLTK toolkit, `https://www.nltk.org/`

Table 4 Most frequent words for each topic for the full corpus, after removing the common frequent words (see text). Words are stemmed and ranked in decreasing frequency. The last row contains the proposed name for each topic.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	featur	agent	predict	control	robot
2	classif	evolv	optim	design	softwar
3	mine	rule	control	evolv	network
4	rule	detect	method	classif	control
5	learn	learn	design	object	method
6	circuit	express	structur	rule	predict
7	network	optim	approach	learn	time
8	method	network	classif	crossov	seri
9	tree	comput	process	featur	evolut
10	evolv	abstract	imag	oper	learn
	Classification	Agents	Optimization	Control	Robots+SW

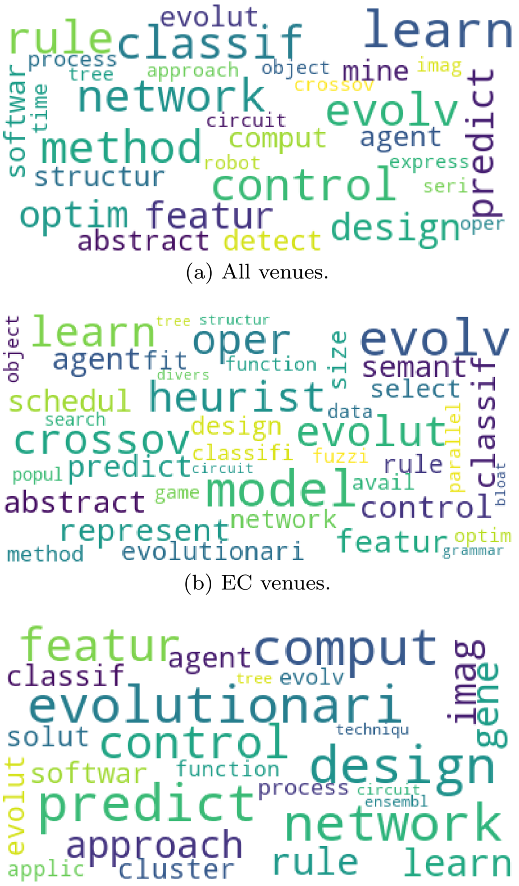
4.2.2 Topics in the full corpus

In this section we examine the full corpus of GP-related documents, i.e., without distinguishing between EC venues and non-EC venues. We executed LDA on the full corpus and noticed that several words were very frequent in several topics (i.e., several topics assigned high probability to those words). For example, “model” turned out to be among the most frequent words for three out of five topics. Other words with high rank in several topics were, e.g., “evolution,” “feature,” “data,” and “function.” While this overlap between sets of highly-ranked words across different topics is not particularly surprising (it confirms that GP literature is mostly about modeling a function based on features of data and using evolution), it also makes more difficult to understand the differences across discovered topics intuitively.

To make the differences among the topics more apparent, thus, we decided to analyze them by removing the most frequent words from the topic descriptions. In detail, for each i-th topic we built the set W_i containing the corresponding 20 most frequent words; then we built the set of the *common frequent words* as $W = \bigcap_{i=1}^{i=5} W_i$. Table 4 describes the topics that emerge from the full corpus after removing all words in W (we omit the description without the removal of those words for brevity). We describe topics with the most frequent words for each topic.

Although it does not emerge any sharp mapping between the resulting topics and established research fields (which on the other hand is a quite common outcome of any literature mining of this kind [12,34,30]), some associations that are broad yet meaningful can indeed be sketched. We summarize those broad associations with a proposed name for each topic as indicated in the last row of the table. Overall, it is interesting to observe that there is no clear separation between topics in terms of theory vs applications. On the contrary, all of the topics exhibit a mix of both aspects (unlike topics that emerge from the corpus composed only of papers published on EC venues, as we shall see in Section 4.2.4).

In Table 5 we show the four most representative documents for each topic (i.e., those for which the association topic-document found by LDA is the strongest). We analyzed those papers in depth and found that most of them are consistent with the high-level description given above. We also found that two documents



(a) All venues.

(b) EC venues.

(c) Non-EC venues.

Fig. 4 Word cloud of the 10 × 5 most frequent words of the five topics for: the whole corpus (4a), the subset of the corpus containing only documents published in EC (4b), and the subset containing only documents in non-EC venues (4c).

associated with the topic named Optimization (the ones highlighted in Table 5) are not actually related to GP. Indeed, those documents are a limitation of our data collection procedure due to the fact documents from other disciplines (e.g., biology) may use the same key terms as GP-related documents (see Section 3). This is not surprising, as GP is a nature-inspired computational technique and, hence, it borrows terms from other nature-related disciplines.

4.2.3 Temporal evolution

We attempted to determine whether the interest of the community for specific topics varied over time. Although LDA does not have these features, we preferred to remain focused on this framework in order to not add further dimensions to our study. An analysis of this kind could be done with a probabilistic topic model in which documents are explicitly associated with a time instant and that may

Table 5 Most representative documents for each topic (see text).

		Document title	Year
Classification	1	Genetic programming applications in chemical sciences and engineering	2015
	2	Accurate and interpretable nanoSAR models from genetic programming-based decision tree construction approaches	2016
	3	Nature-inspired intelligence: A review of selected methods and applications	2009
	4	Implementing linear models in genetic programming	2004
Agents	1	Evolving intelligent systems: Methods, learning, & applications	2006
	2	Evolutionary associative memories through genetic programming	2012
	3	Application of genetic programming for electrical engineering predictive modeling: A review	2015
	4	GenAnneal: Genetically modified Simulated Annealing	2006
Optimization	1	*The cavum septi pellucidi: From embryology to neurosurgery*	2002
	2	*Human obesity: An evolutionary approach to understanding our bulging waistline*	2001
	3	Quantitative electrophilicity measures	2018
	4	A robust data mining approach for formulation of geotechnical engineering systems	2011
Control	1	High performance evolutionary computing	2006
	2	System of systems - From definition to architecture to simulation to space applications	2006
	3	Automatic generation of multipath algorithms in the cellular nonlinear network	2001
	4	Concurrent processing of mixed-integer non-linear programming problems	2009
Robots+SW	1	A web-based water resources simulation and optimization system	2005
	2	Ab initio identification of human microRNAs based on structure motifs	2007
	3	Self-adaptive differential evolutionary extreme learning machines for long-term solar radiation prediction with remotely-sensed MODIS satellite and Reanalysis atmospheric products in solar-rich cities	2018
	4	Robust technical trading strategies using GP for algorithmic portfolio selection	2016

construct a time-variant topic description—i.e., probability distributions across words that may vary over time [6]. We thus considered the LDA model built on the full corpus and associated each document only with its most likely topic. Then, by taking into account the publication date of each document, we counted the number of documents published for each topic and for each year.

Figure 5 shows the absolute number (left) and the percentage (right) of documents for each topic and for each year since the year 2000. It can be seen, in particular from Figure 5 (right), that topic Optimization appears to have increased its popularity over the last 5–10 years, whereas topic Classification and topic Agents exhibit an opposite and clearly decreasing trend.

These figures might suggest a significant trend in the scientific community: while GP-related techniques appear to be increasingly used for problems in the Optimization area, the community appears to be losing interest in using those techniques for attacking problems in the areas related to Classification and Agents. This interpretation could be corroborated by the growing interest in the broad

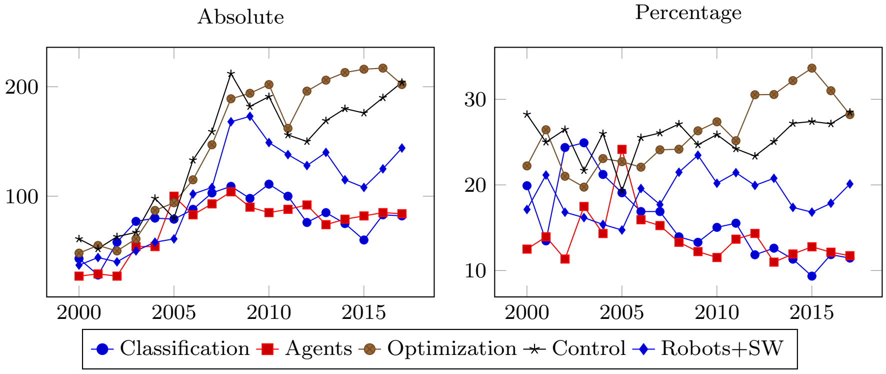


Fig. 5 Number (left) and percentage (right) of documents for each topic (see text) and for each year.

Table 6 Most frequent (stemmed) words for each topic for the corpus containing only documents published in EC venues. Words are stemmed and ranked in decreasing frequency. The last row contains the proposed name for each topic.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	crossov	heurist	heurist	function	model
2	abstract	network	game	semant	learn
3	model	classif	evolv	bloat	evolutionari
4	evolv	data	rule	crossov	circuit
5	design	featur	predict	divers	structur
6	oper	model	grammar	oper	agent
7	evolut	method	evolut	size	represent
8	control	evolv	search	fit	fuzzi
9	avail	learn	schedul	popul	tree
10	parallel	object	classif	select	optim
	EA design	Machine Learning	Discrete opt.	EA dynamics	Applications

area of machine learning and, in particular, of neural networks. We remark again, however, that the actual topics discovered by LSA are much broader than indicated by the names that we have chosen.

4.2.4 Topics in EC and non-EC venues

In this section, we examine GP-related documents published EC in venues separately from those published in non-EC venues. We executed LDA separately on the two corpora and reported the resulting topic descriptions in Table 6 (EC venues) and Table 7 (non-EC venues). The last row of each table provides a proposed name for each topic.

According to Table 6, topics that emerge from the corpus of documents in EC venues are characterized by a significant separation between theory and applications, with two topics that are clearly orientated toward issues related to the mechanics of evolutionary algorithms. The remaining topics are more general but clearly orientated toward the usage of evolutionary algorithms. While these topics are quite broad, the one named Machine Learning appears to be more specific.

Table 7 Most frequent (stemmed) words for each topic for the corpus containing only documents published in non-EC venues. Words are stemmed and ranked in decreasing frequency. The last row contains the proposed name for each topic.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	image	rule	function	softwar	predict
2	predict	classif	control	predict	learn
3	ensembl	evolv	comput	evolutionari	network
4	network	agent	featur	imag	evolutionari
5	comput	network	predict	evolut	gene
6	applic	learn	network	comput	approach
7	design	evolutionari	design	featur	design
8	cluster	tree	evolutionari	control	rule
9	approach	design	circuit	solut	comput
10	techniqu	featur	gene	process	control
	Image	Agents+ML	Control	Software	Networks

On the other hand, topics that emerge from documents in non-EC venues Table 7 are qualitatively more similar to those of the full corpus, in that there is no topic focused on lower-level details of evolutionary algorithms. All the topics are quite broad and not clearly mapped with any specific research field, perhaps with a slightly more explicit correspondence for topics named Agents+ML and Control. This fact could indicate that GP is applied to a broad range of different problems, as well as that there is no specific research field for which GP has become a sort of fundamental tool.

Overall, the topic analysis confirms the initial, relatively straightforward intuition of the authors that can be summarized as follows: EC venues present documents where the main contribution is the improvement of the state-of-the-art performance of GP, while non-EC publication venues exploit these findings to address new challenging problems by means of GP. On the other hand, we expected a somewhat sharper separation between the results from the two corpora and, in particular, a smaller prevalence of applications in EC venues.

We reported in Figure 4 the word clouds of the 10 × 5 most frequent words of the five topics (i.e., the union of the ten most frequent words for each topic) that emerged from the three corpora. The visual inspection of these figures confirms that words related to the mechanics of evolutionary algorithms tend to be more frequent in EC venues than in the other corpora; and, that there is seems to be no strong difference between the full corpus and the one composed only of non-EC venues.

4.3 Citations

This section analyzes the corpus in terms of the number of citations received by each one of the documents.

We remark that the results coming from this analysis should be taken with care. First, it is worthwhile to point out that the number of citations of a scholarly article should not be used as a measure of its quality [4]. Second, our corpus contains very recent documents (up to the current year). In general, a document typically requires some years to reach the entire scientific community and, obviously, the

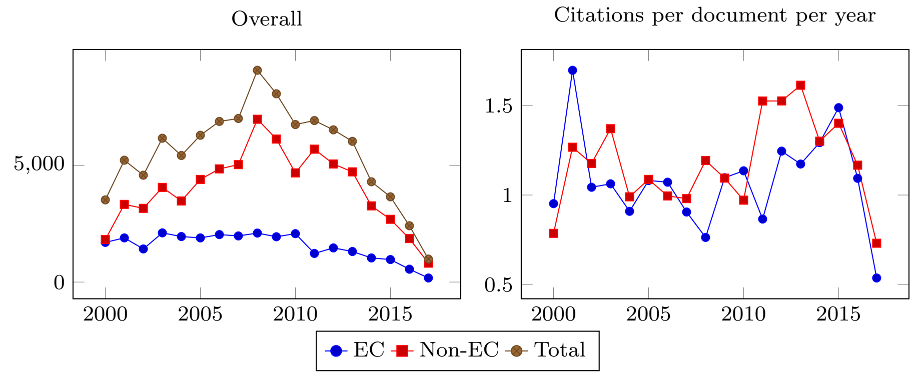


Fig. 6 Overall citations (left) and average citations per document per year (right) received by documents published in EC and non-EC venues.

oldest documents have been available, for the purpose of citation, for longer than the newest ones. As pointed out in [31], the number of citations a document receives is a function of (a) the number of documents published in "close" years (the competing documents) and (b) the number of documents subsequently published (the citing documents) and the number of references they contain.

Despite these limitations, we think that interesting insights may be obtained by analyzing the number of citations.

Figure 6 plots the total number of citations per year (left) and the average citation per document per year (right). We remark that we associate with a year x all of the citations received by documents published in year x: this allows for a comparison between the number of citations of documents published in different years, rather than providing the information about a general tendency to cite GP-related documents by the scientific community.

It can be seen from the right plot of Figure 6 that there is not, in general, an apparent difference in the average number of citations per year received by documents in the EC and non-EC venues. This confirms that studies concerning GP that are published out of EC venues have the same impact of those published in EC venues.

A deeper analysis of the same plot, considered together with the analysis concerning the document type in non-EC venues (Figure 2), seems to reveal that the shift toward journals occurred after 2011 and is resulting in more citations received on average by recent non-EC documents.

4.4 Countries

We wanted to understand which countries most contributed to the GP research field in recent years.

To perform this analysis, for each document we considered the set[8] of the countries inferable from the author affiliations—in case of multiple affiliations of

[8] Originally, countries of affiliation are a multiset as more than one author can be affiliated with an institution in the same country. We considered the corresponding set.

Table 8 Number of documents in our corpus associated with each country.

	Country	Continent	Documents
1	United States	North America	3908
2	China	Asia	3232
3	United Kingdom	Europe	2770
4	Japan	Asia	2456
5	Spain	Europe	1455
6	Iran	Asia	1235
7	India	Asia	1135
8	Canada	North America	1119
9	Italy	Europe	942
10	Brazil	South America	934

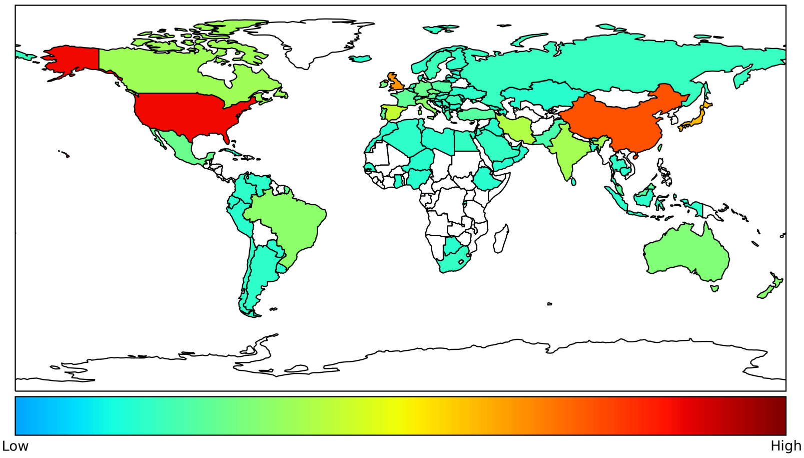


Fig. 7 Contribution of each country to GP documents.

an author, we considered all of the corresponding countries. Then, we associated the document with each country in the set.

Table 8 shows the number of documents in the corpus associated with each of the top ten countries, sorted by number of documents. An overview of the same data for all of the countries can be seen in Figure 7. We remark that these figures are based solely on the number of documents. We do not intend to assess the productivity of the researchers in different countries, which would instead require much deeper analyses based on other data (e.g., population size, funding).

As one can see from Table 8, the United States and China gave the largest contribution to the GP field. This is not surprising as those countries nowadays output the largest fraction of research documents [26]. The United Kingdom immediately follows at third place. Interestingly, two countries from southern Europe are in the top ten list of Table 8, whereas other large European countries (notably,

Table 9 Percentage of documents associated with each country that are associated with each of the topics.

Country	Classification	Agents	Optimization	Control	Robots+SW
United States	16.4	14.7	26.7	26.2	16.0
China	15.3	13.0	30.2	19.1	22.3
United Kingdom	15.4	12.5	23.2	27.6	21.4
Japan	17.4	14.7	16.3	22.7	28.9
Spain	16.8	17.5	22.0	24.9	18.7
Iran	6.2	6.6	56.9	7.7	22.6
India	9.3	13.0	44.5	14.1	19.1
Canada	10.7	13.1	24.2	34.0	18.0
Italy	12.9	15.4	33.2	22.8	15.7
Brazil	22.9	19.7	17.6	25.0	14.8

Germany and France) are not. Raw differences, however, are not that large, as visible in Figure 7.

4.4.1 Topics

For each country, we analyzed the topics associated with the corresponding documents to verify whether the GP research community of that country is more focused on some aspects than others.

Table 9 shows, for each of the top ten countries from Table 8, the percentage of documents associated with each of the five topics. We recall that we associated each document with its most likely topic.

It is evident that the distribution of documents across topics for most countries roughly resembles the distribution obtained by considering all of the documents, with topic Optimization obtaining the largest fraction of the majority of countries. However, it is interesting to note that there are some differences: emerging countries (Iran and India) exhibit a large percentage of documents in topic Optimization; United Kingdom and Canada are, instead, more focused on topic Control.

4.5 Temporal evolution

In order to gain insights into the *temporal evolution* of these data, we aggregated countries in continents and counted the number of documents produced by each continent as a function of time. The results are presented in Figure 8.

It can be seen that Asia and Europe are the continents that contribute the most, both in absolute and relative terms. It can also be seen that the contributions of these two continents are quantitatively very similar since 2010, while before that year there was a relative prevalence of contributions from Europe. Asia exhibited a sort of spike starting in 2007 and peaking in 2009, with values very similar to those of Europe from 2011 onward. In the last two years, though, there has been a non-negligible increase of contributions from Asia with respect to those from Europe.

The contributions from the other continents are, broadly speaking, slowly but steadily increasing, both in absolute and in relative terms. The most significant

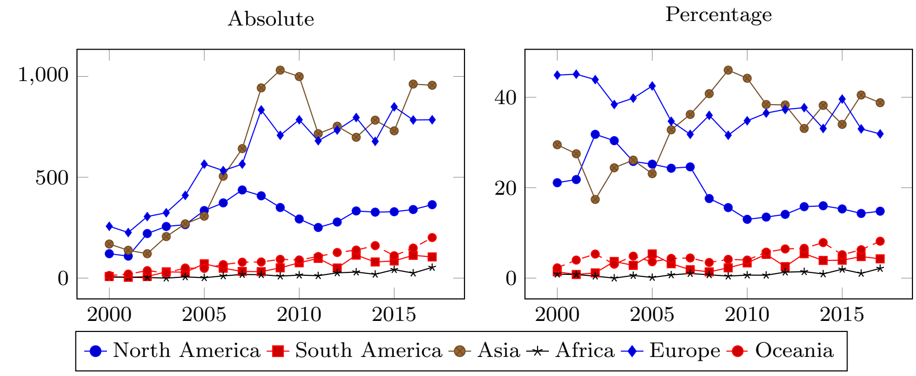


Fig. 8 Number (left) and percentage (right) of documents associated with each continent for each year.

exception to this trend is the drop in the absolute number of documents from North America that occurred in between 2007 and 2010. After 2010, the contributions from North America have started to grow again in absolute terms, albeit more slowly than before, and have remained more or less constants in relative terms.

5 Concluding remarks

We presented an extensive analysis of the GP literature in the 21st century. We focused on the bibliometrics and content of more than 10 000 GP papers. For analyzing the content, we resorted to probabilistic topic modeling, an unsupervised text mining technique that has been used in a broad range of domains. A key element of our study is that we performed our analysis twice: on papers published in venues that are typical of the EC community and on papers published in non-EC venues. This twofold point of view from "both sides of the fence" fostered interesting insights on the overall development and impact of GP as an effective tool for solving problems.

Through our quantitative analyses, we were able to highlight some key trends about the publication habits of our GP community: (1) in recent years, we tend to publish more in non-EC venues; (2) papers published in EC and non-EC venues receive roughly the same number of citations; (3) the research topics that attract the most interest change over the time; (4) emergent countries are joining those that initiated the field in contributing to GP-related research. Despite the intrinsic limitations of a quantitative study of this scale, we believe that our analysis can help to better understand the diffusion, importance, and relevance of GP in the scientific world.

Acknowledgment

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia) under project DSAIPA/DS/0022/2018 (GADgET).

Mauro Castelli acknowledges the financial support from the Slovenian Research Agency (research core funding No. P5-0410).

References

1. The GP bibliography. `http://www.cs.bham.ac.uk/~wbl/biblio/`. Accessed: Oct 2018
2. Statistics & Collaboration Network in GECCO. `https://www.researchgate.net/publication/318701234_GECCO_Statistics_Collaboration_Network`. DOI 10.13140/RG.2.2.25153.66404
3. Bao, G., Fang, H., Chen, L., Wan, Y., Xu, F., Yang, Q., Zhang, L.: Soft robotics: Academic insights and perspectives through bibliometric analysis. Soft robotics **5**(3), 229–241 (2018)
4. Bartoli, A., Medvet, E.: Bibliometric evaluation of researchers in the internet age. The Information Society **30**(5), 349–354 (2014)
5. Blei, D.M.: Probabilistic topic models. Communications of the ACM **55**(4), 77–84 (2012)
6. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: ICML (2006)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
8. Branke, J., Nguyen, S., Pickardt, C.W., Zhang, M.: Automated design of production scheduling heuristics: A review. IEEE Transactions on Evolutionary Computation **20**(1), 110–124 (2016)
9. Dabhi, V.K., Chaudhary, S.: Empirical modeling using genetic programming: a survey of issues and approaches. Natural Computing **14**(2), 303–330 (2015)
10. De Nart, D., Degl'Innocenti, D., Pavan, A., Basaldella, M., Tasso, C.: Modelling the user modelling community (and other communities as well). In: International Conference on User Modeling, Adaptation, and Personalization, pp. 357–363. Springer (2015)
11. Espejo, P.G., Ventura, S., Herrera, F.: A survey on the application of genetic programming to classification. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **40**(2), 121–144 (2010)
12. Evangelopoulos, N., Zhang, X., Prybutok, V.R.: Latent semantic analysis: five methodological recommendations. European Journal of Information Systems **21**(1), 70–86 (2012)
13. Harzing, A.W., Alakangas, S.: Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. Scientometrics **106**(2), 787–804 (2016)
14. Herrera, M., Roberts, D.C., Gulbahce, N.: Mapping the evolution of scientific fields. PloS one **5**(5), e10,355 (2010)
15. Koza, J.R.: Survey of genetic algorithms and genetic programming. In: WESCON/'95. Conference record.'Microelectronics Communications Technology Producing Quality Products Mobile and Portable Power Emerging Technologies', p. 589. IEEE (1995)
16. Langdon, W.B., Gustafson, S.M.: Genetic programming and evolvable machines: ten years of reviews. Genetic Programming and Evolvable Machines **11**(3-4), 321–338 (2010)
17. McDermott, J., White, D.R., Luke, S., Manzoni, L., Castelli, M., Vanneschi, L., Jaskowski, W., Krawiec, K., Harper, R., De Jong, K., O'Reilly, U.M.: Genetic programming needs better benchmarks. In: Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation, GECCO '12, pp. 791–798. ACM (2012)
18. McKay, R.I., Hoai, N.X., Whigham, P.A., Shan, Y., O'Neill, M.: Grammar-based genetic programming: a survey. Genetic Programming and Evolvable Machines **11**(3), 365–396 (2010)
19. Medvet, E., Bartoli, A., Davanzo, G., De Lorenzo, A.: Automatic face annotation in news images by mining the web. In: Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01, pp. 47–54. IEEE Computer Society (2011)
20. Medvet, E., Bartoli, A., Piccinin, G.: Publication venue recommendation based on paper abstract. In: Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on, pp. 1004–1010. IEEE (2014)
21. Meguebli, Y., Kacimi, M., Doan, B.L., Popineau, F.: Unsupervised approach for identifying users' political orientations. In: European Conference on Information Retrieval, pp. 507–512. Springer (2014)
22. Mongeon, P., Paul-Hus, A.: The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics **106**(1), 213–228 (2016)

23. Oltean, M., Groşan, C., Dioşan, L., Mihăilă, C.: Genetic programming with linear representation: a survey. International Journal on Artificial Intelligence Tools **18**(02), 197–238 (2009)
24. Petke, J., Haraldsson, S., Harman, M., White, D., Woodward, J., et al.: Genetic improvement of software: a comprehensive survey. IEEE Transactions on Evolutionary Computation (2017)
25. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on Web search and data mining, pp. 399–408. ACM (2015)
26. Schlegel, F., Schneegans, S., Eröcal, D.: UNESCO science report: towards 2030. UNESCO Publ. (2015)
27. Sondhi, P.: Feature construction methods: a survey. Tech. rep. (2009)
28. Tremblay, M.C., Parra, C., Castellanos, A.: Analyzing corporate social responsibility reports using unsupervised and supervised text data mining. In: International Conference on Design Science Research in Information Systems, pp. 439–446. Springer (2015)
29. Vanneschi, L., Castelli, M., Silva, S.: A survey of semantic methods in genetic programming. Genetic Programming and Evolvable Machines **15**(2), 195–214 (2014)
30. Velden, T., Boyack, K.W., Gläser, J., Koopman, R., Scharnhorst, A., Wang, S.: Comparison of topic extraction approaches and their results. Scientometrics **111**(2), 1169–1221 (2017)
31. Wallace, M.L., Larivière, V., Gingras, Y.: Modeling a century of citation distributions. Journal of Informetrics **3**(4), 296–303 (2009)
32. White, D.R., McDermott, J., Castelli, M., Manzoni, L., Goldman, B.W., Kronberger, G., Jaśkowski, W., O'Reilly, U.M., Luke, S.: Better gp benchmarks: community survey results and proposals. Genetic Programming and Evolvable Machines **14**(1), 3–29 (2013)
33. Xue, B., Zhang, M., Browne, W.N., Yao, X.: A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation **20**(4), 606–626 (2016)
34. Yau, C.K., Porter, A., Newman, N., Suominen, A.: Clustering scientific documents with topic modeling. Scientometrics **100**(3), 767–786 (2014)