A 50 year retrospective on academic integrity and computer ethics in CS Education

SIGCSE Special Projects - Final Project Report
May 2019

Nadia Najjar, Mary Lou Maher, Maryam Mohseni
{ nanajjar, mmaher9, mmohseni}@uncc.edu
University of North Carolina at Charlotte
Department of Software and Information Systems

Abstract

In recent years, the topics of academic integrity and computer ethics have surfaced frequently in the computer science education community as academics continue to find ways to deal with cheating and plagiarism in their higher education courses. In this project, we investigate how academic integrity and teaching computer ethics has evolved within the SIGCSE community over the past 50 years. We apply Natural Language Processing techniques, including frequency analysis of bag of words and probabilistic topic modeling, to understand how these topics have emerged within the community to maintain academic integrity and help develop ethical future computer professionals. We present the labels for the topics, word frequencies, and a temporal visualization of the frequency of occurrence of specific topics.

1. Introduction

In this project, we seek to investigate how academic integrity and teaching computer ethics has evolved within the SIGCSE community over the past 50 years. “Integrity is at the core of all effectively functioning societies and organizations” [1]. Fostering an environment that promotes integrity is especially important in academia, a “self governing society where open discussion and democracy should prevail in all aspects of teaching, curriculum, and research” [1, 2]. The Internet has dramatically revolutionized the way information is transferred around the globe. This shift in information transfer has enabled a world-wide access of course materials and assignment solutions to everyone, making academic and professional integrity a challenge in every field of study. This is especially the case within Computer Science where students understand how technology can potentially be used to plagiarize and copy/paste code from online sources.

The ACM’s Code of Ethics is built on three main principles: trust, respect and privacy. It is defined as the “collection of principles and guidelines designed to help computing professionals make ethically responsible decisions in professional practice. It translates broad ethical principles into concrete statements about professional conduct” [3]. Training students to think ethically and preparing them to be active moral agents is critical in today’s world.

Over the past 50 years, a subset of SIGCSE papers have explored the principles of academic integrity and computer ethics, and shared best practices of successful mechanisms that instructors can use in their

classes to address both. In this paper, we apply Natural Language Processing techniques to highlight publications and trends in two areas 1) ethics, and 2) academic integrity.

In this paper, we describe the corpus of 50 years of SIGCSE Symposia data and develop various time-series visualizations to identify trends and patterns on the topics of academic integrity and computer ethics. We apply Natural Language Processing techniques, including word frequencies on a bag of words representation of titles and abstract, and topic modeling on paper abstracts.

2. Description of the MetaData from 50 Years of the SIGCSE Symposium Series

For this project SIGCSE provided us with the dataset of the proceedings of the SIGCSE Technical Symposium starting from the first technical symposium in 1970 and ending with the 49th Technical Symposium in 2018. The dataset included a metadata JSON file with information about each record in the proceedings.

The metadata file consists of 6207 records, where each record is defined as a dictionary of key-value pairs. Each key refers to an attribute of the record, and its value shows the value of that attribute for that specific record. There are a total of 40 distinct keys/attributes among all records in the file. Each record consists of multiple attributes and their corresponding values but not all records contain all of these 40 attributes (some records miss some of these attributes). One of these attributes is called **“recordType"** which has the value of “article” for all the 6207 records in the metadata (JSON) file. Therefore all records are considered as an article; indeed each record gives information about a distinct article (after this in this report we may call a record as an article and a key as an attribute as these can be used interchangeably). The bar chart in Figure 1 shows 40 distinct attributes we found in the metadata file with the count of articles (records) containing that attribute.

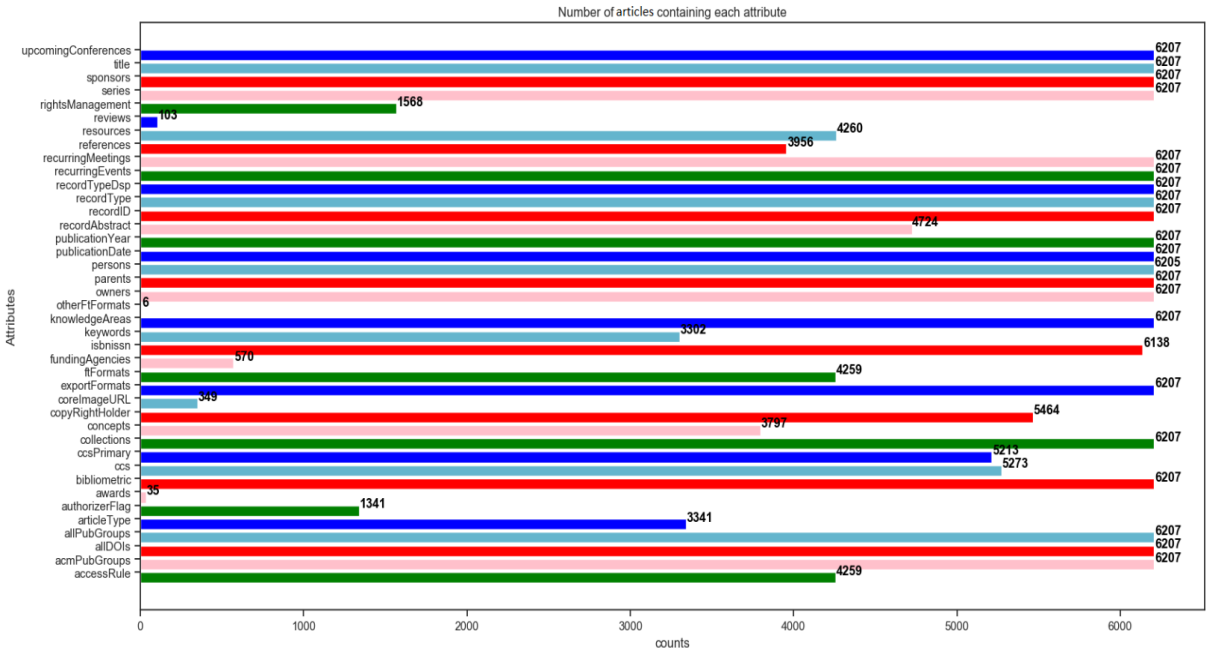


Figure 1. Number of articles (records) containing each of 40 attributes

"articleType": is one of the attributes provided for just 3341 of the records (articles). We found there are a total of 14 different article types as shown by the chart in Figure 2 (note that "articleType" is different from "recordType". As discussed, "recordType" is set to "article" for all the 6207 records but "articleType" is provided for only 3341 of the records/articles). 1211 of the articles are labeled as "Regular Article". We found that these articles labeled as "Regular Article" are related to the papers published mostly between 2008 and 2018. The abstract for each corresponding article (that is scientific paper in the case for regular articles) is provided in the "recordAbstract" attribute of the record. We will explore and analyze the regular articles separately and in more detail in section 2. The other article types are named as: "Demonstration" as labeled to 40 articles, "Extended Abstract" labeled to 15 articles, "Forum" 12 articles, "Invited Talk" 3 articles, "Keynote" 36 articles, "Panel" 131 articles, "Poster" 370 articles, "Section" 743 articles, "Short Paper" 13 articles, "Technical Note" 115 articles, and "Tutorial" labeled to 103 articles. 2866 of the records/articles are not provided with the "articleType" attribute which are mostly related to the articles for before 2008. We call the type of these articles "unknown" and explore them in section 3.

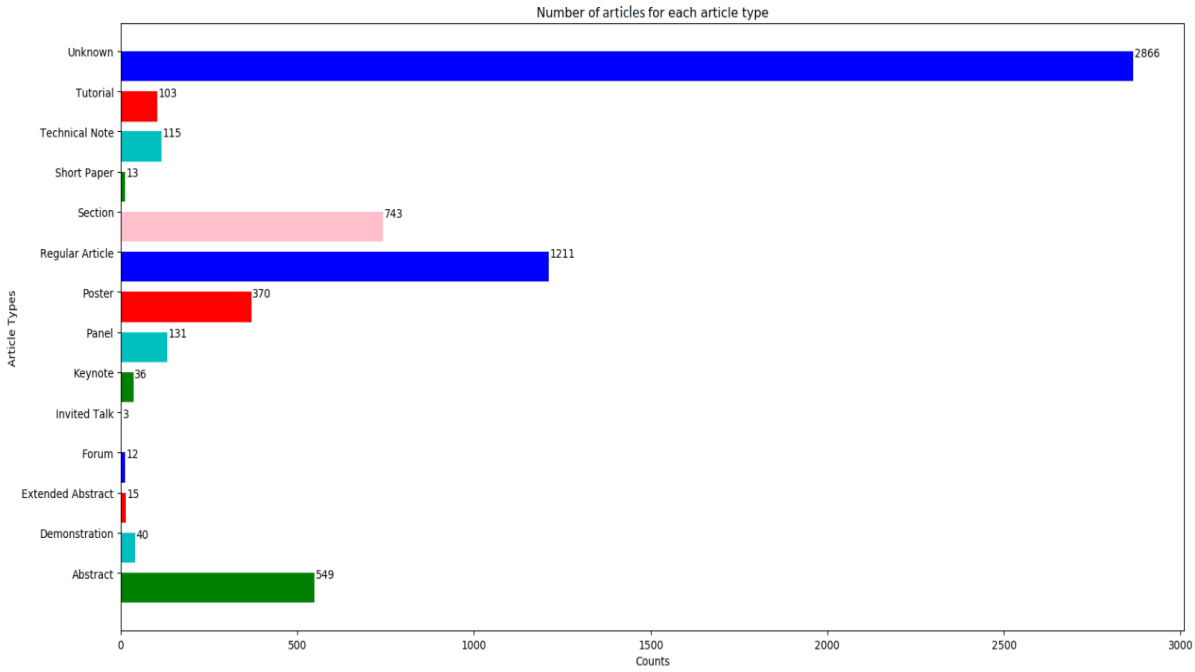


Figure 2. Number of articles for each article type

"publicationYear": is another article attribute provided for all articles. The publication year for the articles in the metadata file (regardless of their article type) range between 1970 and 2018. We will analyze the publication years for regular articles in section 2.

"knowledgeAreas": is another attribute provided for all of the 6207 articles. It consists of three sub-attributes: "knowledgeAreaID" ,"knowledgeAreaAcronym", and "knowledgeAreaURL". By analyzing the record for all articles we found that there are 2 different types of "knowledgeAreaAcronym": 1)"Education" assigned to 6080 of the articles, and 2)"Interaction" assigned to 127 of the articles. This attribute does not distinguish the papers as the "Knowledge Area" comes from the ACM assigned Knowledge Area for SIGCSE (Education and Interaction).

"concepts": The other attribute is "concepts" provided for 3797 of the articles (not all 6207). It consists of three sub-attributes: "concept", "conceptAbstract" (providing a description of the concept), and "conceptURI". By analyzing the records in the metadata, we found that there are a total of **856** different concepts among all records/articles. As it is not practical to list all of 856 concepts or show all of them by bar charts, we visualize them using word clouds (see Figure 3). Also, the 10 most frequent concepts with the counts of articles having those concept types are shown in the bar chart in Figure 4. As shown in Figure 4, "Abstraction (computer science)" is the most frequent concept assigned to 177 articles as their concept. "Computer Program" with 164 articles is the second most frequent concept, etc. Figure 3 and Figure 4 show the concepts for all existent articles (records) in the metadata regardless of their article types. We will inspect the concepts for regular articles in section 4.4.

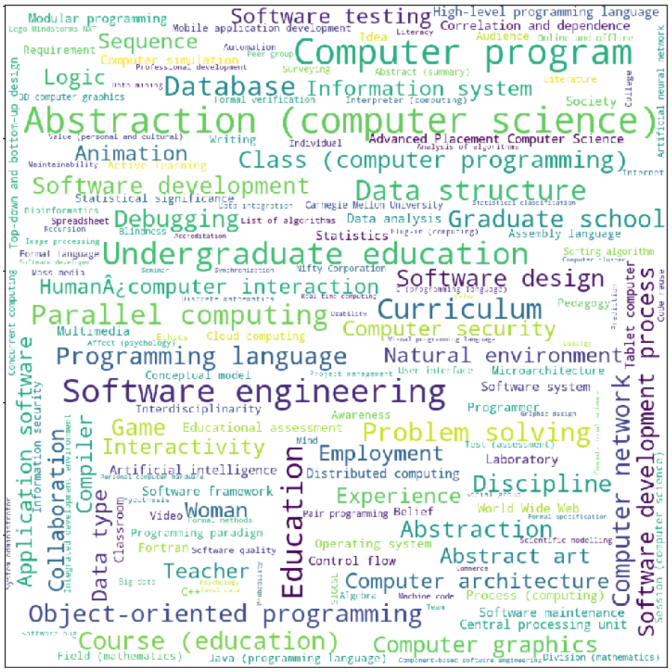


Figure 3. Word cloud of different concepts for all articles

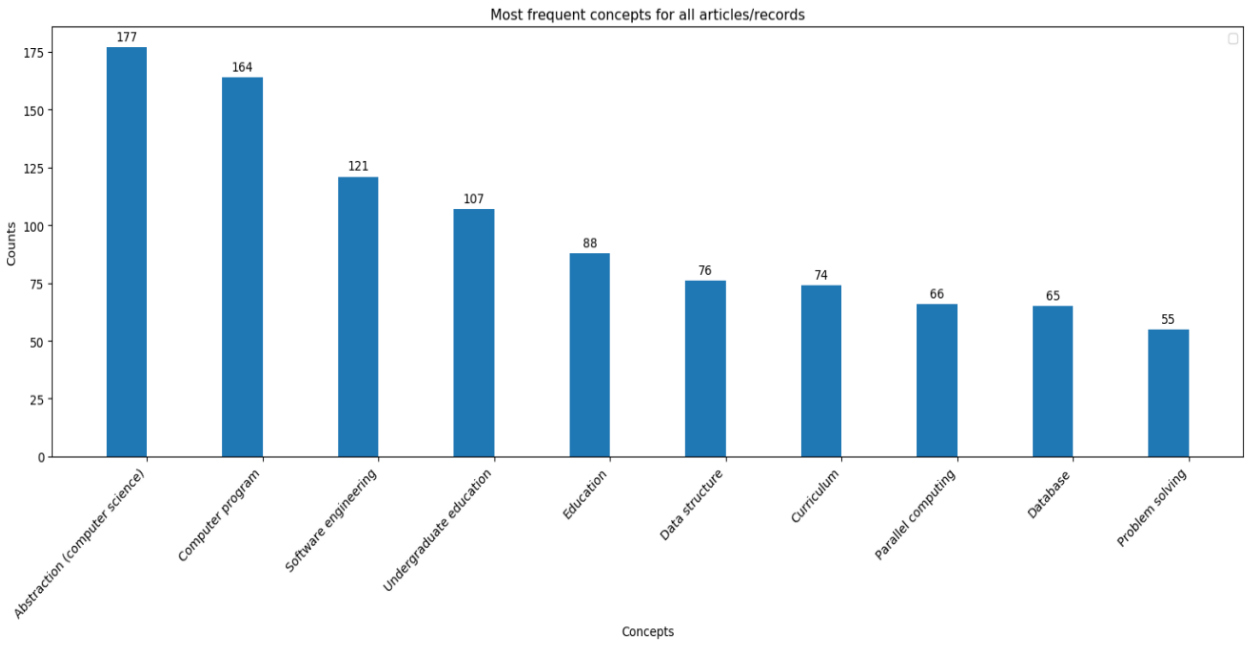


Figure 4. Ten most frequent concepts among all articles (records) with number of articles

"recordAbstract": this attribute provides the abstract of the published scientific paper for regular articles, or a brief description of the article for other article types (e.g. some information about the conference, section, etc.). Only 4724 out of 6207 articles have the "recordAbstract" attribute.

“persons”: is another attribute provided for all of the 6207 articles/records. It consists of different sub-attributes describing the information of the article’s authors like author’s name, affiliation, etc.

The rest of the attributes are mostly self explanatory from their name. In overall, the main attributes we may deal with among all the 40 attributes are: “recordID”, “title”, “articleType”, “publicationYear”, “recordAbstract”, “concepts”, and “keywords” (containing the authors defined keywords).

Analysis of Metadata for articles labeled as “Regular Articles”:

As stated earlier, “articleType” is one of the attributes provided for 3341 of the articles. 1211 articles out of these 3341 ones have the articleType of “Regular Article”. By inspecting the regular articles, we found that they are regular scientific papers published in proceedings of SIGCSE. In this section we analyze the main attributes of these articles which are labeled as “Regular Article” by SIGCSE.

“publicationYear”: The articles labeled as “Regular Article” by SIGCSE are published between 2008 and 2018. The bar chart in Figure 5 displays the count of published regular articles for each year. As can be seen, the number of regular articles published in 2018 is considerably more than the other years. We assume that may be because of an extension in acceptance rate of submitted papers to SIGCSE for the year 2018.

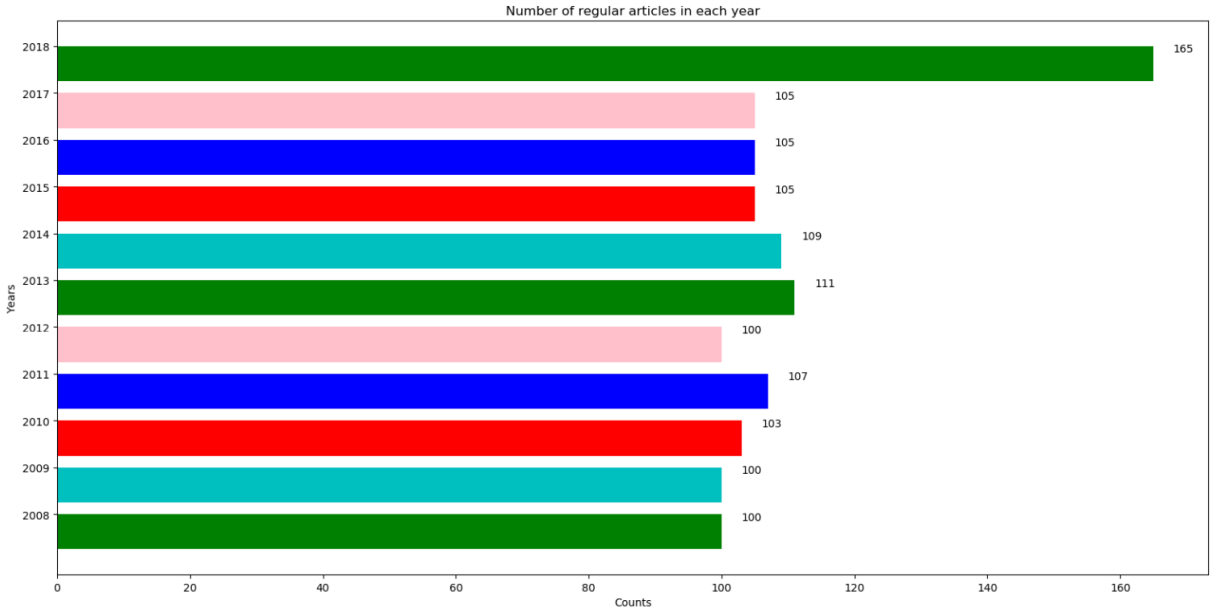


Figure 5. Number of regular articles published in each year

“knowledgeAreas”: 1111 of the 1211 regular articles have the “knowledgeAreas” attribute of “Education”. The remaining 100 regular articles have the “knowledgeAreas” attribute of “Interaction”.

“concepts”: We found that among 1211 regular articles, only 706 have the “concepts” attribute and the remaining 505 regular articles do not have the attribute of “concepts”. By analyzing these 706 regular

articles which contain the “concepts” attribute, we obtained a total of 317 different concepts. Figure 6 below displays the word clouds containing these concepts. Ten most frequent concepts for regular articles are shown in the bar charts in Figure 7. As illustrated in Figures 6 and 7, the most frequent concept for regular articles is “Software engineering” assigned to 23 of the regular articles. “Computer program” and “Education” are the second and third frequent concepts assigned to 22 and 18 regular articles respectively.

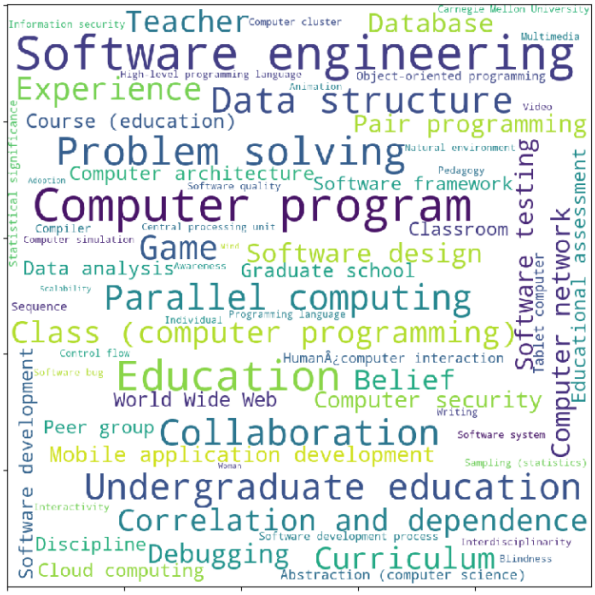


Figure 6. Word cloud of different concepts for regular articles

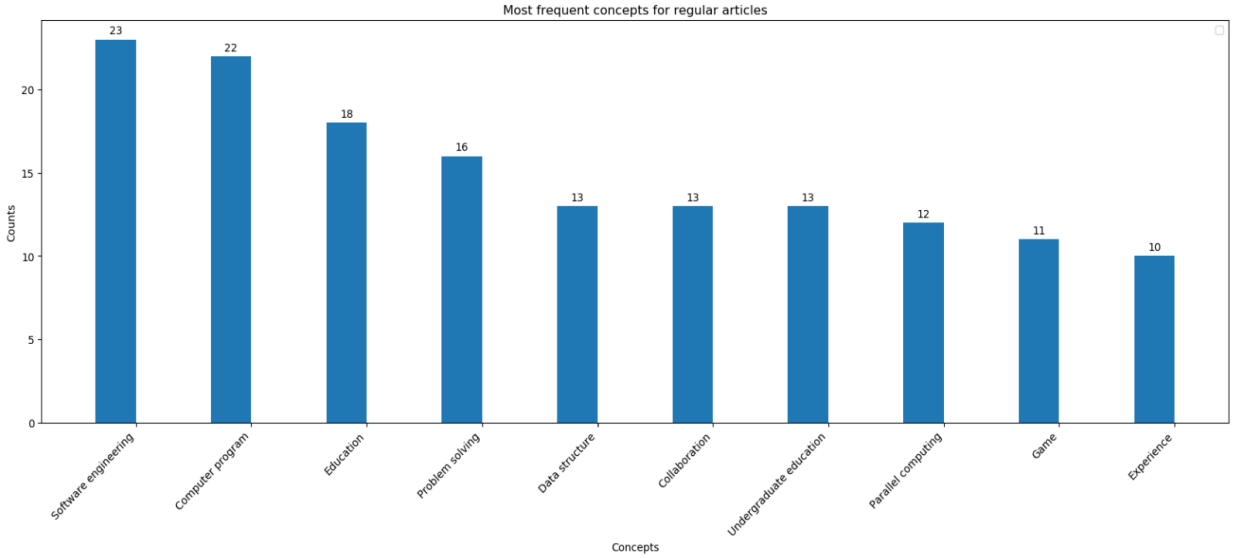


Figure 7. Ten most frequent concepts for regular articles with number of regular articles having them

3. Natural Language Processing Models

To understand the trends and patterns that emerged within the topics of ethics and academic integrity within the SIGCSE community we used a Natural Language Processing (NLP) to identify documents that are relevant to the two themes we are interested in. We start with one of the most fundamental NLP algorithms, Bag-of-words (BOW) [5]. BOW can extract the important features from unstructured text. BOW is used as a pre-processing step where the text is converted into a large corpus of words and the number of their occurrences.

The unstructured text data in the dataset that can be used as the source for the BOW algorithm are the titles for each document and abstract text for a subset of the documents. Converting this text to a BOW representation documents that include words that are relevant to the themes used in this analysis can be identified.

We extracted the titles for the 6207 documents available in the metadata file and created a bag-of-words model for each document and a model for the complete corpus where we included all the documents. We will refer to this model as the *title-based BOW* model.

The metadata file included abstracts for 4724 documents. Similar to the approach we followed using the text from the titles we created a bag-of-words model using the abstract text. We then explored the model for words related to ethics and academic integrity. We will refer to this model as *abstract-based BOW* model.

The second NLP approach we applied was Topic Modeling [7]. We identified 1211 records labeled as "Regular Article" in the dataset and also had a value for the abstract attribute that can be used as the text to represent the documents. Topic modeling is a probabilistic model for learning the themes that occur in a collection of documents [7]. These models produce a set of "topics" each comprising a distribution over all the words in the corpus. This is based on the modeling assumption that topics are a probabilistic mixture of all the words in the corpus. Words which feature strongly in a topic gain a relatively high probability. Each document in the corpus is assigned with different proportions of each topic as in a mixture model. For example, a paper may be drawn from 40% a topic and 60% from a combination of other topics. Topics are not labeled by the system and they are not promised to consist of a single theme that is easily human understandable, but they are usually at least moderately interpretable.

4. Trends and Patterns in SIGCSE Proceedings

The two main themes we are interested in are academic integrity and teaching computer ethics.

For each theme we started by identifying a list of relevant subtopics we were interested in exploring and examining how they emerged within those two main themes over the past 50 years within the SIGCSE community.

The subtopics for the academic integrity theme are: integrity, cheating, morals, rules of conduct, wrongdoing, plagiarism, falsification. The ethics theme subtopics are, minorities, legal issues, under represented groups, inclusivity, diversity, discrimination, and morals.

4.1 Coverage

To understand the coverage of the two themes and their subtopics within the SIGCSE community we extracted the documents where one of the words of the theme or the subtopic appeared at least once.

Using the titles of the documents we found 113 from the 6207 documents related to the ethics theme and subtopics which is less than 2% of the documents while only 26 (0.4%) documents were relevant to the academic integrity theme. Using the abstracts text the number of documents for each theme was very close. The ethics theme had 294 documents which is close to 6% and 95 documents which is 2% in the academic integrity theme.

Tables [1] show the breakdown of the number of documents for each of the themes and their perspective subtopics.

Theme	**Title-based Model**	**Abstract-based Model**
Ethics	113	294
Academic Integrity	26	95

Table 1. Number of documents in each theme identified with BOW model

4.2 Theme-level Overlap

Using the set of documents for each of the themes we found two documents that appeared in both themes generated from the analysis based on document titles. Table 2 lists those two documents.

Year	Article Type	Title
2009	Regular Article	A model academic ethics and integrity policy for computer science departments
2017	Abstract	The ACM Code of Ethics and Professional Conduct: Teaching Strategies and the Coming Update

Table 2. Documents with overlapping themes based on titles

Examining the documents identified based on the abstracts there were 16 documents that appeared in both themes. Figure 8 shows the distribution of the document types in this subset. As mentioned in Section 2, documents published in 2007 or earlier did not have an entry for the article type attribute. There were 5 documents with an “unknown” article type. We were able to manually identify 4 of those documents by referencing the PDF files. This chart shows that the two themes appeared in various formats within the SIGCSE community with the majority of them in the “Regular Article” format.

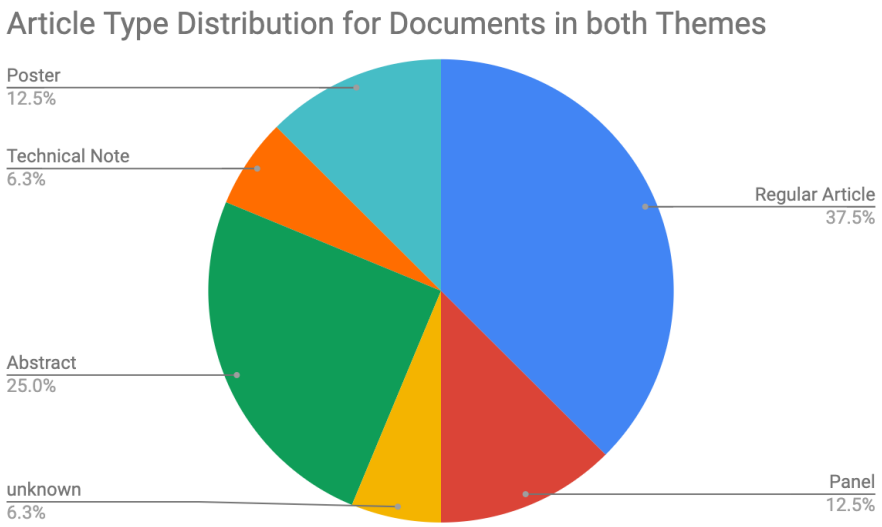


Figure 8. Distribution of the different types of documents with overlapping themes using abstract-based model

Figure 9 shows the chronological distribution of the documents in this subset as they appeared over the 50 years of SIGCSE. A listing of the documents identified in this subset can be found in Appendix A.

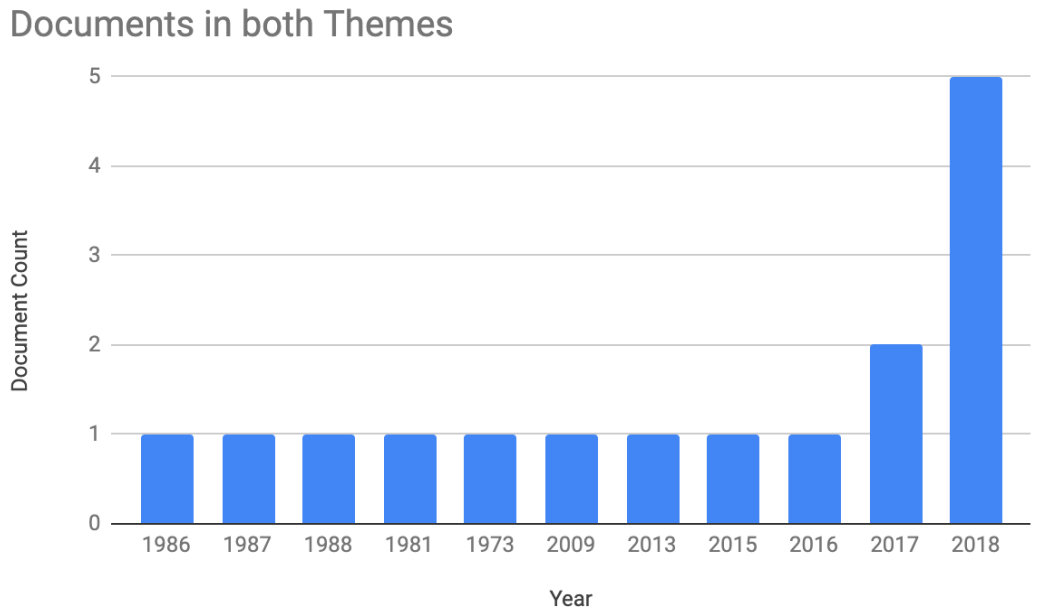


Figure 9. Distribution of overlapping documents using abstract-based model

3.4 Subtopic-level Overlap

To understand the relationship within each of the themes and their subtopics we identified the document overlap that exists within each theme. We identified the 113 documents relevant to the ethics theme and its subtopics using the title-based model. Table 3 shows the breakdown of those documents within this subset. We notice here that the majority of the documents appeared under the ethic and diversity subtopics and there were zero documents identified under the discrimination or morals subtopics.

Ethics Subtopics	# of documents
Ethic	48
Minorities	6
Legal issues	3
Under represented groups	14
Inclusivity (inclusive)	5
Diversity	40
Discrimination	0
Morals	0

Table 3. Document count for Ethics theme subtopics using title-based model

In this subset three documents had overlap containing words from two subtopics within the ethics theme. The subtopic overlap was between ethics and legal; inclusivity and diversity; and minority and underrepresented. The majority of documents with a known type in this subset belonged to the "Regular Article" type. Figure 10 and Table 4 show the distribution of the document types in this subset.

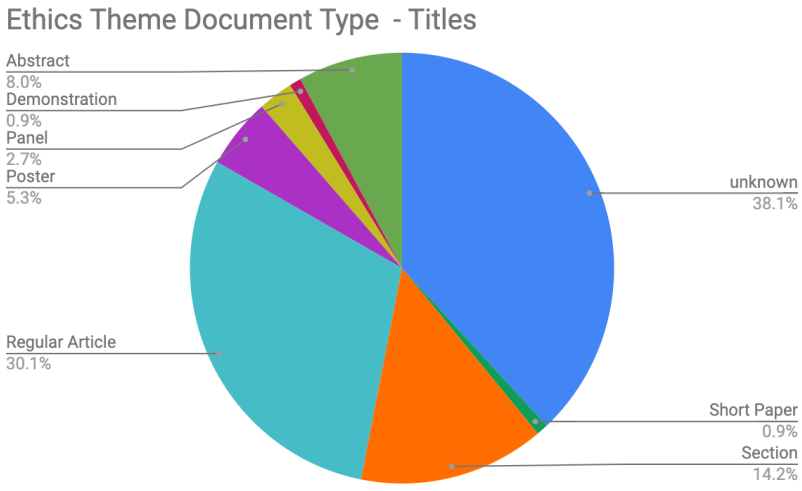


Figure 10. Distribution of document types in Ethics theme using the title-based model

Document Type	Count
Unknown	43
Tutorial	0
Technical Note	0
Short Paper	1
Section	16
Regular Article	34
Poster	6
Panel	3
Keynote	0
Forum	0
Extended Abstract	0
Demonstration	1
Abstract	9

Table 4. Count of document types in Ethics theme using the title-based model

Applying the abstract-based model to the Ethics theme we found 294 relevant documents. Table 5 shows the number of documents in each of the subtopics of this theme. We expected the model based on the abstracts to result in more documents identified as relevant since abstracts have more text than titles.

Ethics Subtopics	# of documents
Ethic	73
Minorities	47
Legal issues	7
Under represented groups	99
Inclusivity (inclusive)	17
Diversity	88
Discrimination	4
Morals	0

Table 5. Document count for Ethics theme subtopics using abstract-based model

Similarly more overlap appeared between the subtopics of the theme. From this set of 294 documents, 51 documents appeared in 2 or more subtopics with the highest number of document overlap (20 documents) between the minority and underrepresented subtopics. Table 6 shows the number of documents shared between the subtopics.

Subtopics	**# of documents**
Ethic, legal	7
Ethic, diversity	1
Ethic, discrimination	1
Underrepresented, inclusive	5
Minority, underrepresented	20
Underrepresented, diversity	11
Minority, inclusive	2
Minority, diversity	4
Minority, underrepresented, discrimination	2

Table 6 Subtopics with overlapping documents in Ethics theme using the abstract-based model

The majority of documents in this subset belonged to the "Regular Article" type similar to the title-based model. Figure 11 and Table 7 show the distribution of the document types in this subset. This model shows the ethics theme appearing in different types of documents from what the title-based model revealed. Documents with types of tutorials, technical notes, keynotes and extended abstracts appear in this theme.

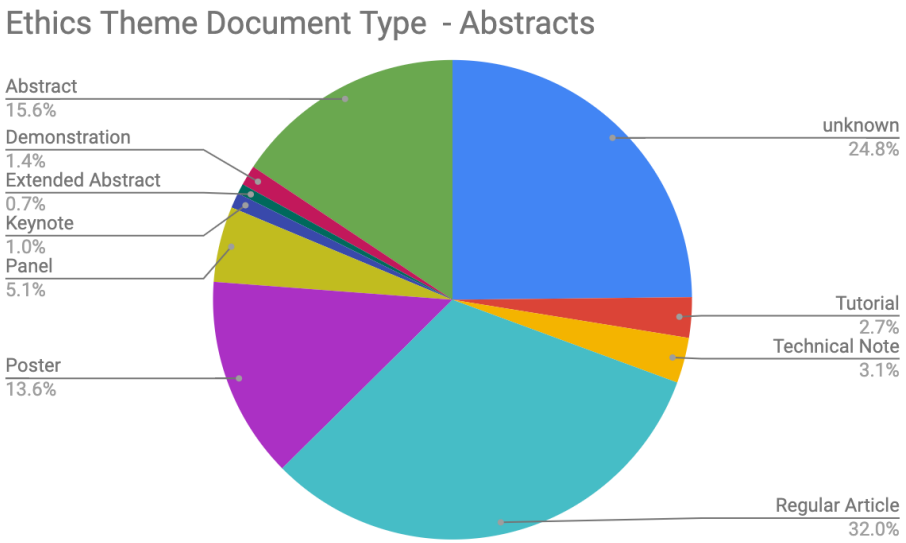


Figure 11. Distribution of document types in Ethics theme using the abstract-based model

Document Type	Count
Unknown	73
Tutorial	8
Technical Note	9
Short Paper	0
Section	0
Regular Article	94
Poster	40
Panel	15
Keynote	3
Forum	0
Extended Abstract	2
Demonstration	4
Abstract	46

Table 7. Count of document types in Ethics theme using the title-based model

Next we explored the documents in the academic integrity theme to understand the relationship within its subtopics. Similar to our approach with the ethics theme we explored the results using both the title-based and abstract-based models.

We identified the 26 documents relevant to the academic integrity theme and its subtopics based on the title-based model. Table 8 shows the breakdown of those documents within this subset. We notice here that only 3 documents appeared under the integrity subtopic while the majority of documents appeared under the plagiarism subtopics and conduct subtopics. There were zero documents identified under the remaining subtopics. There were no documents that overlapped between the subtopics in this theme.

Academic Integrity Subtopics	# of documents
integrity	3
Plagiarism	14
Cheating	0
Morals	0
conduct	9
Wrongdoing	0
Falsification	0

Table 8. Document count for Academic Integrity theme subtopics using title-based model

Figure 12 and Table 9 show the distribution of the document types in this subset. This model shows the academic integrity theme appearing in different types of documents. Not considering documents with “unknown” types, this subset’s documents appear mostly in “abstract” format.

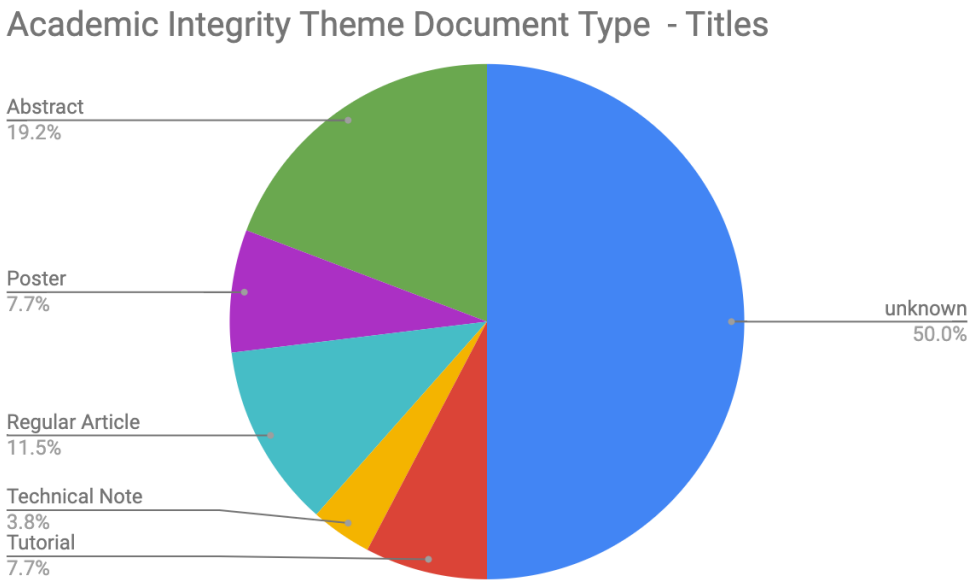


Figure 12. Distribution of document types in Academic Integrity theme using the title-based model

Document Type	Count
Unknown	13
Tutorial	2
Technical Note	1
Short Paper	0
Section	0
Regular Article	3
Poster	2
Panel	0
Keynote	0
Forum	0
Extended Abstract	0
Demonstration	0
Abstract	5

Table 9. Count of document types in Academic Integrity theme using the title-based model

Using the abstract-based model we identified 96 documents belonging to the academic integrity theme. Table 10 shows the breakdown of these documents within this subset. This table shows that the vast majority of the documents were relevant to the conduct theme with 45 documents and that wrongdoing and falsification subtopics had no relevant documents.

Academic Integrity Subtopics	# of documents
integrity	20
Plagiarism	23
Cheating	15
Morals	3
conduct	45
Wrongdoing	0
Falsification	0

Table 10. Document count for Academic Integrity theme subtopics using abstract-based model

Figure 13 and Table 11 show the distribution of the document types in this subset. This model shows a similar coverage as the title-based model for this theme with respect to the documents type.

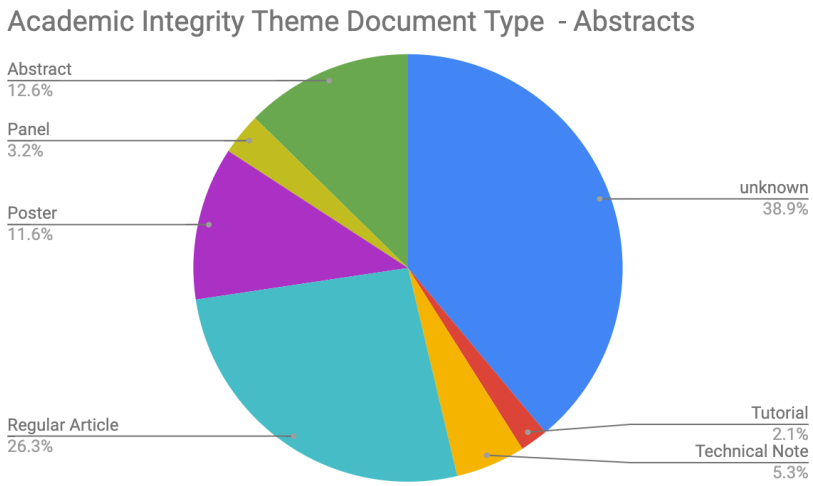


Figure 13. Distribution of document types in Academic Integrity theme using the abstract-based model

Document Type	Count
Unknown	37
Tutorial	2
Technical Note	5
Short Paper	0
Section	0
Regular Article	25
Poster	11
Panel	3
Keynote	0
Forum	0
Extended Abstract	0
Demonstration	0
Abstract	12

Table 11. Count of document types in Academic Integrity theme using the title-based model

Taking a closer look at the documents in each of the subtopics we found 8 documents overlapping between 2 or 3 subtopics. Half of these documents appeared under the plagiarism and cheating subtopics. Table 12 shows the breakdown of these documents and the subtopics.

Subtopics	**# of documents**
Plagiarism, cheating	4
integrity, cheating	1
integrity, conduct	1
cheating, conduct	1
Integrity, plagiarism, conduct	1

Table 12 Subtopics with overlapping documents in Academic Integrity theme using the abstract-based model

4.3 Temporal Analysis

To understand the emergence of the ethics theme in the titles for 50 years of SIGCSE Technical Symposium proceedings we plot the chronological distribution of this subset of documents in Figure 14. This information shows that the Ethics theme first emerged in the titles of the documents in 1975. There was one document belonging to this theme in that year with the title "Innovative computer services for minority colleges" [4]. Given the date of this document the data did not include the classification for what type of document it was and we were not able to manually label it. To get an understanding of what type of context this document appeared within the community we looked at other information in the metadata such as the abstract text and determined it was a paper related to equity and minorities.

Figure 14 also shows a noticeable increase in the number of documents in recent years starting in 2005 with 2018 seeing the highest number of the 50 year

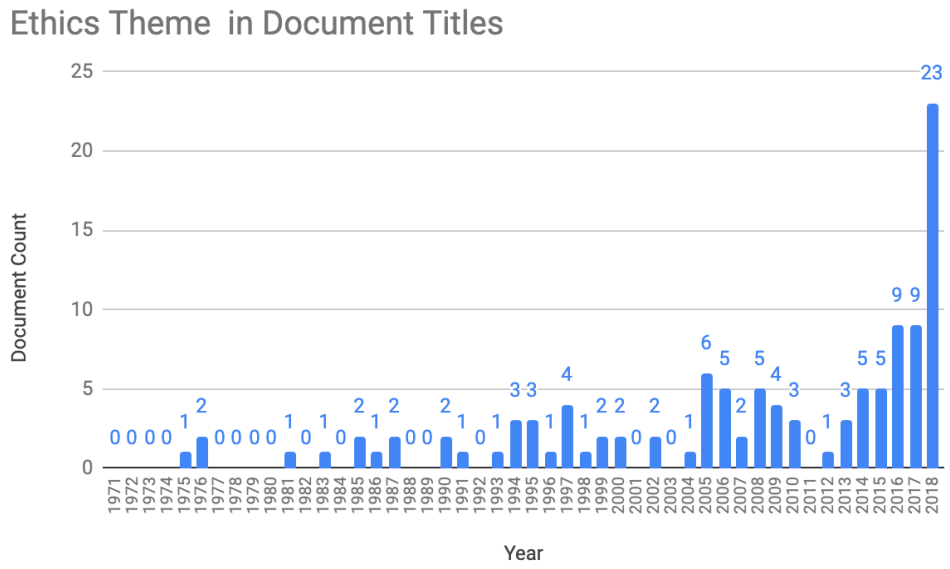


Figure 14 - Chronological distribution of documents in the Ethics theme from the title-based model

To understand the chronological pattern in this theme using the abstract-based model we plot the distribution of this subset of documents in Figure 15 over the 50 years. This plot shows an overall similar trend as the title-based model. With one document appearing for the first time in 1973. This document was titled "On the structure of a computing profession" [6]. Reviewing the abstract of this paper we discern that it is related to integrity and ethical consideration in computer systems. The plot also shows the increasing interest in the ethics theme starting in 2005 with 2018 seeing the highest number of documents related to this theme. Appendix B shows a temporal distribution of the documents in each subtopic.

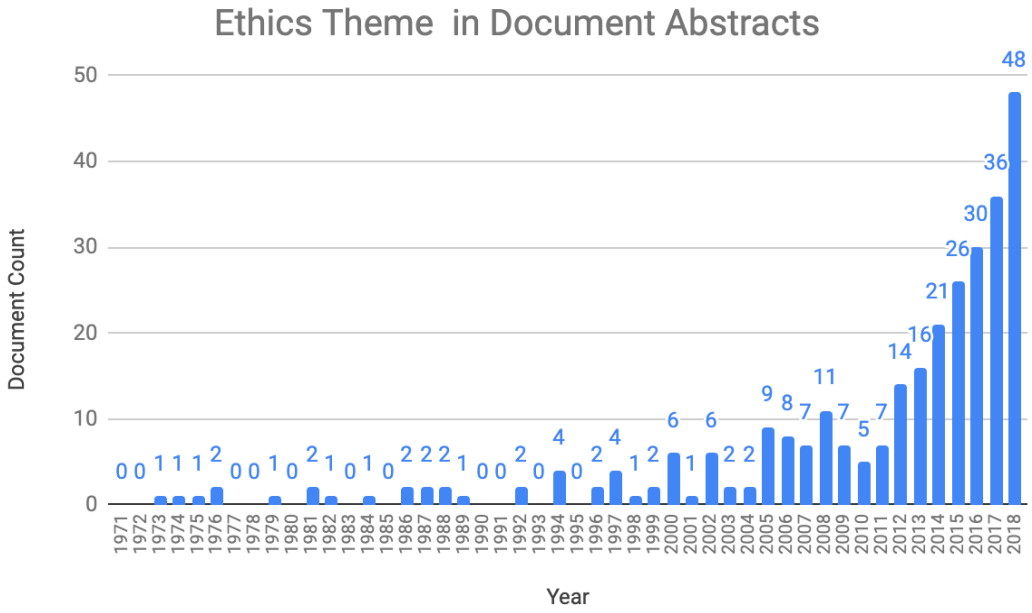


Figure 15 - Chronological distribution of documents in the Ethics theme from the abstract-based model

We conduct a similar analysis to understand the emergence of the academic integrity theme.
Figure 16 shows the chronological distribution of the documents identified using the title-based model for this theme; we notice that this theme had a slow start. The first document appeared in 1978 followed by a few documents scattered in the 1980s. Documents started appearing again in the early 2000s mostly after 2005 and 2017 saw the highest number with 6 documents in one year.

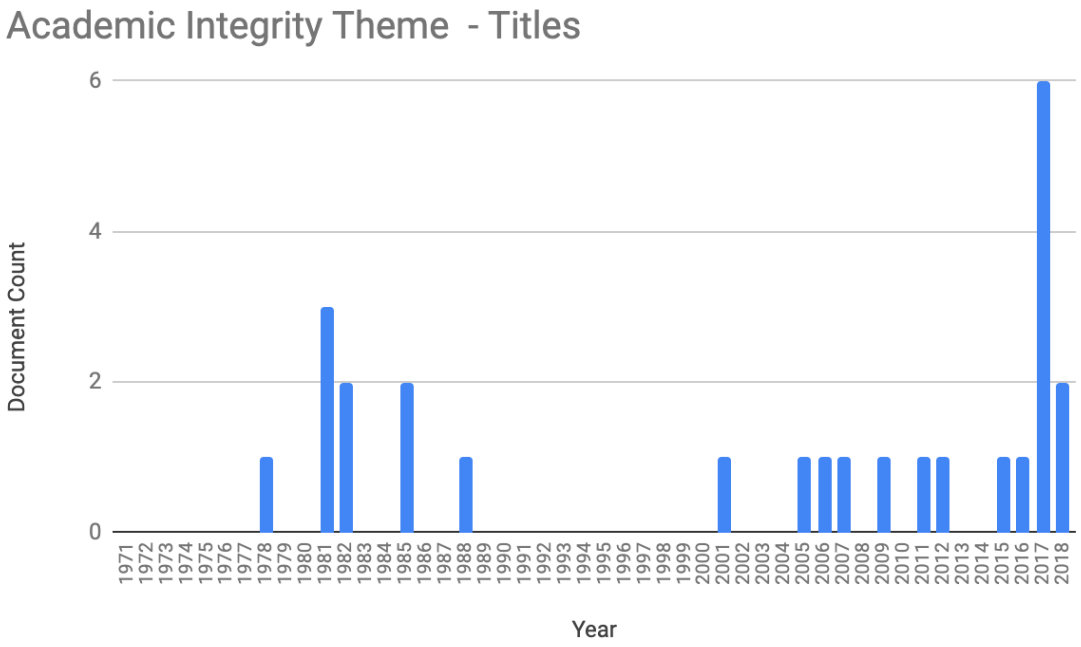


Figure 16 - Chronological distribution of documents in the Academic Integrity theme from the title-base model

The abstract-based model shows an earlier emergence of documents in this topic as shown in Figure 17. Similar to the ethics theme both BOW models show an overall similar trend in the academic integrity theme. The most interesting observation is the increasing interest in this theme since 2005 and spiking in the last two years. Appendix C shows a temporal distribution of the documents in each subtopic.

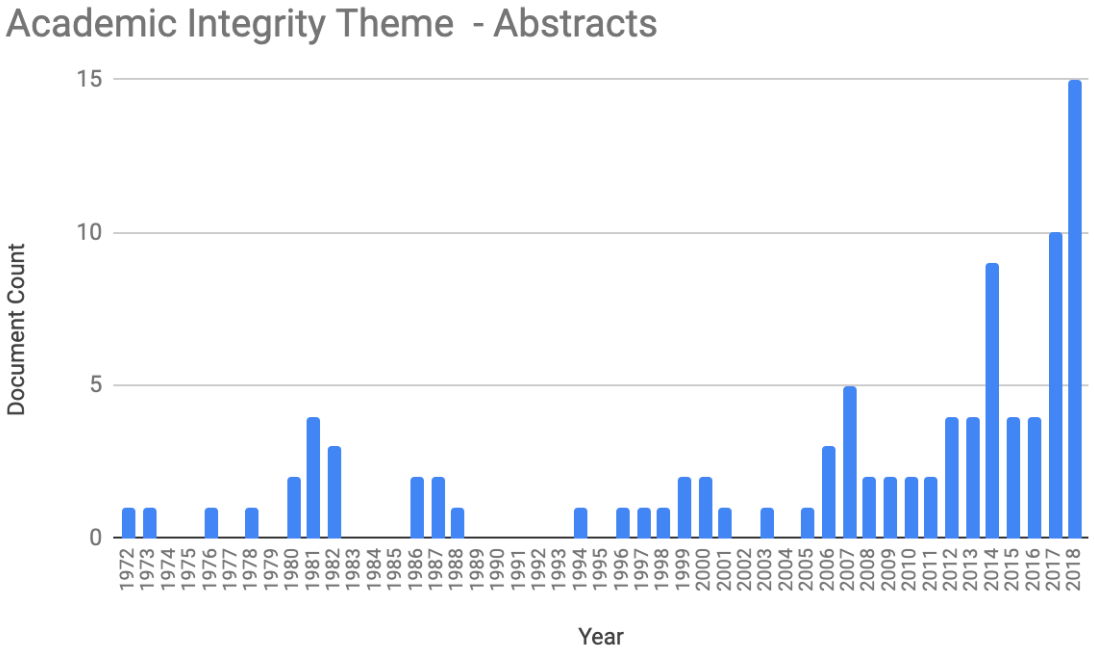


Figure 17 - Chronological distribution of documents in the Academic Integrity theme from the abstract-base model

4.4 Topic Modeling of "Regular Articles"

We extracted the abstracts of articles labeled as "Regular Article" by SIGCSE (1211 articles) from the metadata file to run topic modeling on them. As discussed earlier, "recordAbstract" attribute of the regular articles contains the abstract of the published article. We created a dataset from the abstracts of the regular articles including recordID, publication year, title and the abstract itself. We run topic modelling on the corpus of the regular article abstracts to derive the main topics (themes) of regular articles [7]. We used the R package "STM" to run topic modeling on the dataset of the regular article abstracts. STM (Structural Topic Model) [8] is a topic model extension that is equivalent to CTM (Correlated Topic Model) [9] implementation in some configurations. In order to prepare the data to feed to the STM algorithm, we removed the stopwords, numbers, and punctuations by applying appropriate functions included in the STM package. All the words were converted to lowercase and then stemming was performed. After preprocessing steps, STM topic model algorithm was run on the dataset of the abstracts. We used the default number of topics, 20, in our study. By running topic modeling we represented each document in the corpus with a 20-dimensional vector containing the distribution of topics in that document. We used the abstract of the papers provided in the metadata as a summary of each paper.

Table 13 shows the most representative words for each of the 20 topics. Some of the words seem not to be the complete word (such as "simul" in Topic 12) as they are the stem of the combination of several words with the same root. e.g. "simulate", "simulation", and "simulator". Singular word forms are combined with their plurals in stemming. Most of the obtained topics have understandable meanings within the domain of Computer Science Education research. Our interpretation about the subject of each topic is included in the last column of Table 13.

Table 13: Top most representative words for 20 topics and our interpretation of the overall subject for each topic

Topic #	Most representative words	Our interpretation
Topic 1	pair-programming, error, program, trace, message, student, novice, industry-inspired	Pair-programming, industry-inspired, programming
Topic 2	assignment, submission, homework, feedback, review, grade, student, program	Assignment, feedback, assessment
Topic 3	parallel, computing, robot, multicore, shared-memory, microlab, hardware, course, program, student, architecture	Parallel-computing, multicore, architecture
Topic 4	system, query,database, machin, sql, python, program, paper, tool, language, virtual, web, user-level	Database, python, web-application
Topic 5	Flipped-course, peer, learn, class, active-learning, instructor, face--face, lecture, adjunct, rate, online, student	Flipped-course, active-learning, peer-learning
Topic 6	teacher, school, workshop, profession, reform, comput, develop, high,train, CSP (Computer Science Principles), middle-school	school, training, professional-development

Topic 7	self-confid, gender, femal, retent, male, comput, differ, tutor, prior,assist, student	Gender, retention, tutor
Topic 8	algorithm, problem, solve, recurs, approach, misconcept,proof, induction, student	Algorithm, recursion, induction
Topic 9	assessment, program, comput, attitud, mentor,camp , women, intervention,commiss, accredit	Mentor, women, intervention
Topic 10	secure, software, develop, industry, agile, team, pedagogical code reviews, Test-Driven Development, student, engin	Security, code-reviews, agile
Topic 11	major, minor, undergraduate, biology, comput, bioinformat, underrepresentation, recruit, enrolment, women, student, program	Underrepresentation, minor, recruit
Topic 12	test, simul, implement, framework, processor, datapath, machin, cpu, function, control, run, unit, version, student, program	Simulation, testing, frameworks
Topic 13	data, visual, structur, big data, collect, graph,cluster, algorithm, analytics, tool	Big-data, visual-analytics
Topic 14	game, video, cybersecur, player, student, learn, project-based learning, design, team, competit, process oriented guided inquiry learning (POGIL)	Game, project-based-learning, cybersecurity
Topic 15	program, languag, concept, comput, environ, scratch, kodu, Alice, children, animation,student	Programming-languages, programming-environments, children
Topic 16	ethic,think, computational creativity exercises (CCEs), comput, cours, education, metaphor, policies	Ethics, creativity
Topic 17	mobil, network, devic, system, kernel, laboratori, student, cours, project, comput, platform	Mobile, network, kernel
Topic 18	exam, question, student, perform, learn, grade, predict, score, final, correl, geek, wrapper	Predictive-models, student-performance
Topic 19	social, tour, music, earsketch, remix, culture, collabor, comput, student, learn, design, engag	Music, social-learning, engagement
Topic 20	code, write, object-ori, program, student, cours, introductori, comput, image, read, media	Media-computation, object-oriented, introductory-computing

For example, Topic 1 seems to be mostly about programming, pair-programming and industry-inspired matters. Topic 2 is about assessment and feedback. Topic 3 is clearly about parallel-computing and architecture, while Topic 5 is clearly about flipped courses, active learning, and peer-learning. Some topics include words that are related to the language used to explain research itself, rather than a very

specific subject like Topics 6 and 8. Topic 13 clearly refers to a particular research subject, big data and visual analytics. Our interpretation for other topics can be found in Table 13. We observed some meaningless words in some topics like "csp" in Topic 6. We searched our dataset for these kinds of meaningless words and found that they are abbreviations for some phrases; that was "*Computer Science Principles*" for "csp". So we included the actual phrases in parentheses next to the abbreviations in Table 13. These topics reasonably reflect the abstracts of our dataset.

Among these 20 topics identified by topic modeling, Topic 11 (underrepresentation, minor) and Topic 16 (ethics) related to the two topic themes we are interested in. We found that 144 (12%) of this set of regular articles have Topic 11 as their first or second highest proportion and 159 (13%) of regular articles have topic 16 as their first or second highest proportion. These results are inline with what the BOW models showed in terms of the emerging for these themes over the years. The documents that appear in this set were all published in the last 10 years (2008 - 2018) and these percentages show the importance of these themes in these recent years. Topic modeling produces a topic correlation matrix showing the relationship between any two topics. Topic 11 was most correlated with Topic 16 and had a slightly higher correlation than other topic pairs.

4.5 Topic Modeling of Documents Discovered by BOW Model

We applied topic modelling to the 439 records that were identified by the BOW analysis as relevant to the themes of ethics and academic integrity. We set the number of topics 10 since this was a smaller set of documents. The results of this model are displayed in Table 14 showing the most representative words for each of the 10 topics as well as our interpretation of the main topic. This approach identified the two main themes of interest in this project in topic 2, ethics, and topic 6, academic integrity. It also shows topic 9 related to minority and underrepresentation which is one of the subtopics under the ethics theme.

To see how much the topic modeling and BOW models overlap we found 304 records showing Topic 2, Topic 6, or Topic 9 as one of their top two topic proportions. We note here that this subset of documents included different types of documents, not just "Regular Article" documents. This meant that the abstract text varied in quality and the type of information it contained. The abstracts of regular research papers are typically summaries of the papers which can encompass the topic of the paper. This is not necessarily the case when the article type is not "regular articles" such as: "poster", "section", "tutorial", "technical notes", "Panel", "Keynote", "demonstration", and "forum".

Table 14: Top most representative words for 10 topics and our interpretation of the overall subject for each topic for records identified by BOW on ethics and academic integrity

Topic #	Most representative Words	Our interpretation
1	workshop, develop, csp (Computer Science Principles), curriculum, teacher, comput, school, cours, learn, audit, classroom-readi	Computer science in schools
2	ethic, social, issu, legal, privaci, paper, cours, confront, copyright, intellectu, patent, law, property, topic	Ethics, legal issues
3	project, team, work, member, faculti, process, learn, experi, peer, strategi, department, tide, cope, supervis	Project-based learning, teamwork, peer learning
4	game, exercis, lab, simul, class, engag, learn, use, experi, approach, teach, physic	Gamification
5	pair, perform, signific, exam, score, cours, studi, learn, result, anova, mental, post-test, pre-test, measure, ict (Information and Communication Technology)	Experimental studies
6	plagiar, homework, submiss, cheat, grade, solut, difficulti, program, assign, method, problem, multiple-choic, quizz, suspect, submiss, detect	Plagiarism, cheating (academic integrity)
7	graduat, job, industri, busi, bachelor, access, program, educ, system, technolog, hire, textil, commerc, conduct	Industry jobs for graduates
8	languag, program, project, java, experi, messag, trace, novic, learn, cobol, declar, encapsul, error, compil, recurs, ide	Programming languages
9	women, underrepres, femal, rise, career, major, minor, girl, camp, comput, school, program, particip, allianc, asian, high-qual, sister	Women, underrepresented, minority
10	research, intervent, studi, formal, primari, conduct, educ, pedagogi, learn, paper, venu, studio, psychology, experi	Research interventions

5. Summary

This analysis highlighted two important topics that have seen an increase in interest over the years in the years within the SIGCSE community specifically and higher education in general: ethics and academic integrity. Our analysis shows how this interest emerged in the context of other important topics such as diversity and underrepresentation. Our temporal analysis shows

that it is not until 2007 or 2008 that we begin to see high word frequencies in the SIGCSE proceedings related to ethics and academic integrity. The academic integrity and ethics concerns are aligned with the widespread use of the internet and availability of online resources that have been highly associated with academic integrity violations in higher education courses. We found that only 12% of the regular articles since 2008 are highly proportionally correlated with the topics related to academic integrity and 13% with ethics. These results show that the trend is an increasing focus on ethics and academic integrity, but only in recent years.

Acknowledgment

This project was supported by a Special Award from the SIGCSE Special Projects Committee.

References

[1] Calabrese, R, and Roberts, B. 2004. Self-interest and scholarly publication: the dilemma of researchers, reviewers, and editors. *International Journal of Educational Management*. 18(6)

[2] Gerber, L. 2001. I nextricably linked: shared governance and academic freedom. *Academe.* 87(3).

[3] ACM Code of Ethics and Professional Conduct. 2018. Retrieved from https://www.acm.org/code-of-ethics

[4] Jesse C. Lewis. 1975. Innovative computer services for minority colleges. SIGCSE Bull. 7, 1 (February 1975), 7–10. DOI:https://doi.org/10.1145/953064.811122

[5] Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. Springer.

[6] Mills, H. D. 1973. On the structure of a computing profession. SIGCSE Bull. 5, 1 (February 1973), 97–101. DOI:https://doi.org/10.1145/953053.808087

[7] Blei D. M., Ng A., and Jordan M. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research. 3 (4–5): pp. 993–1022. 2003.

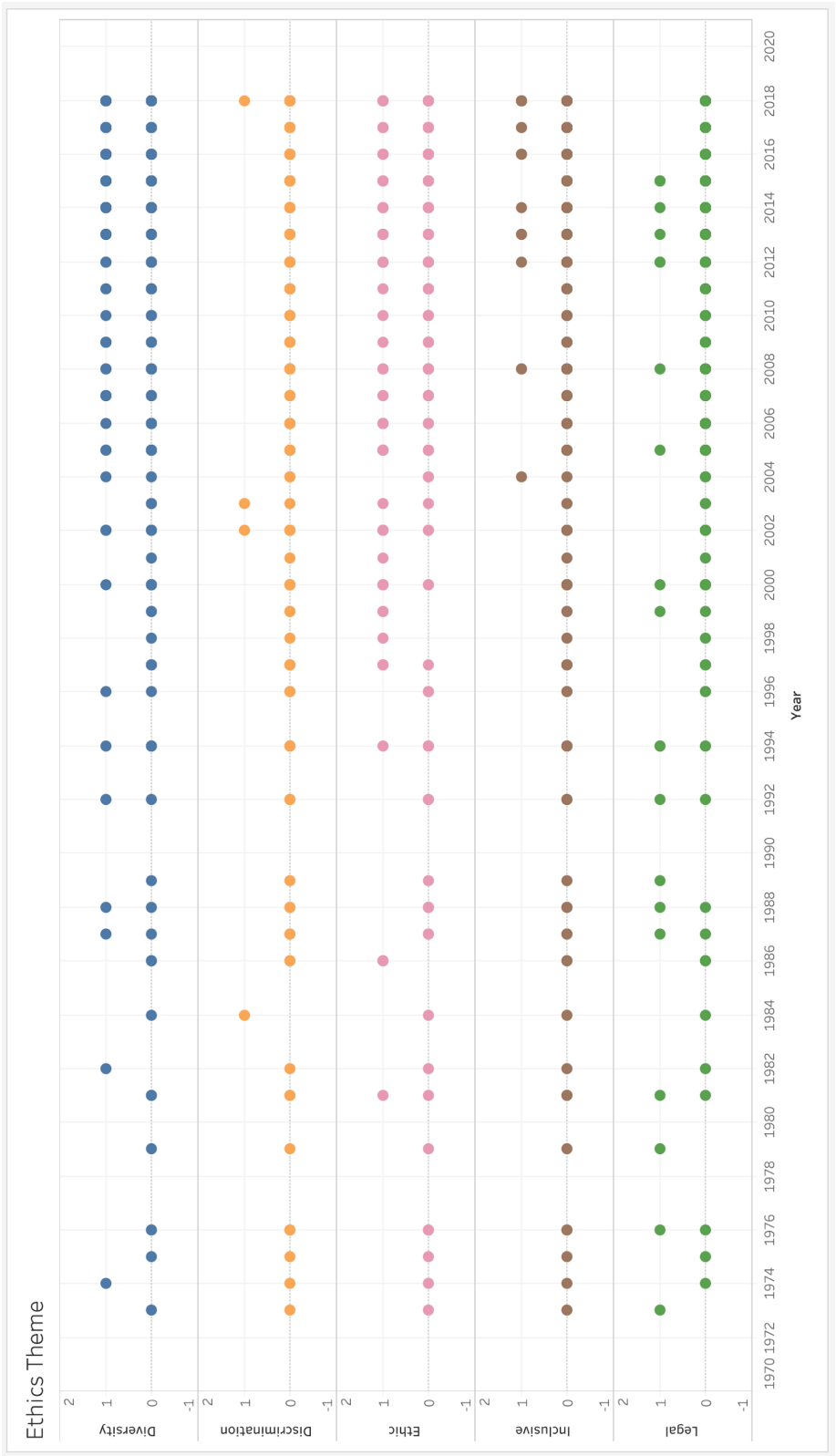
[8] Roberts M., Stewart B., Tingley D., Lucas C., Leder-Luis J., Gadarian S., Albertson B., and Rand D. 2014. Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science. 58(4), 1064-1082.

[9] Blei D. M., and Lafferty J. D. 2007. A correlated topic model of Science. The Annals of Applied Statistics 1.1, 17-35 (2007). doi:10.1214/07-AOAS114.

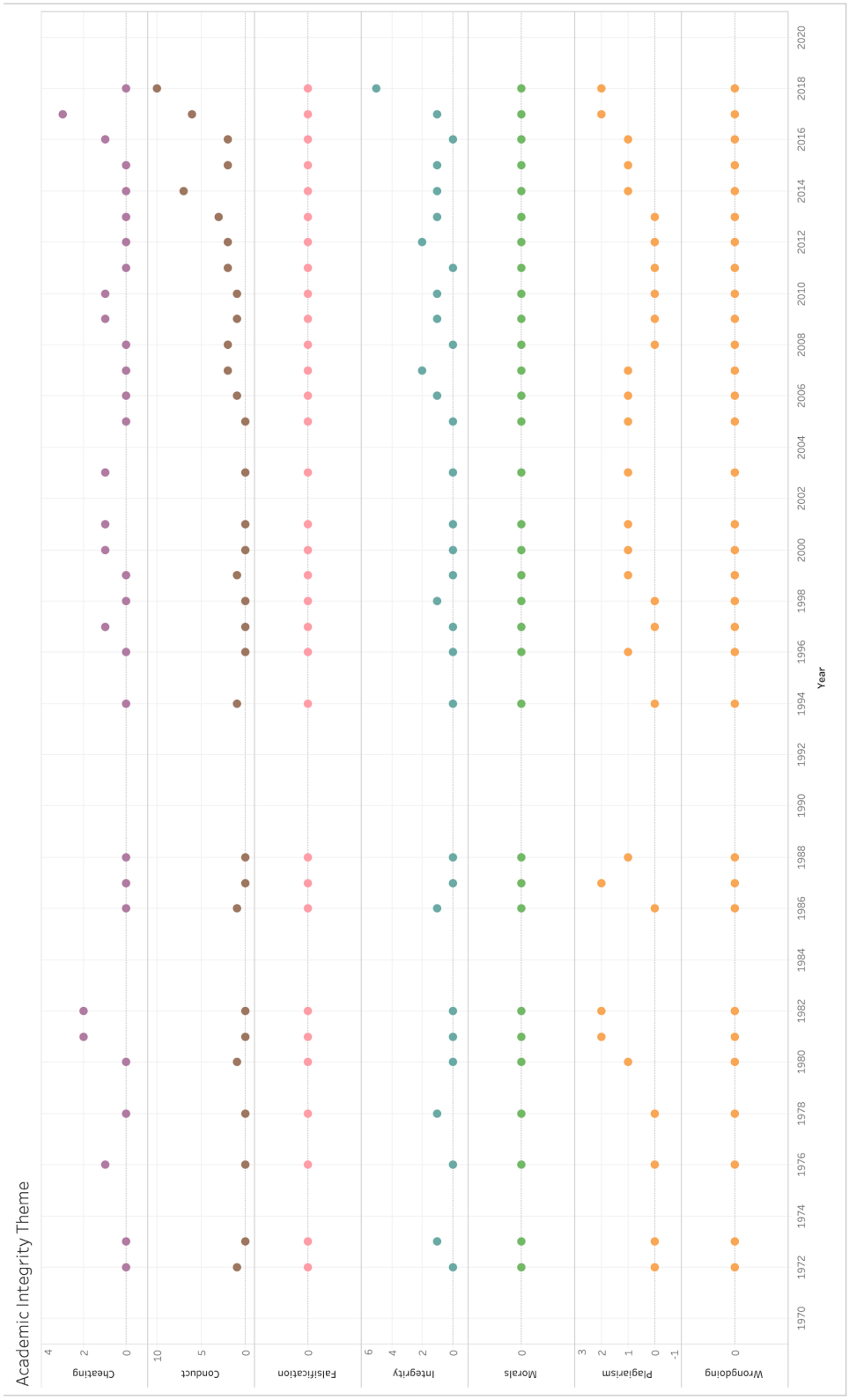
Appendix A

Year	Type	Title
1986	unknown	Increasing students security awareness: article II. What C.S. graduates don`t learn about security concepts and ethical standards
1987	unknown	Defining ethical and unethical student behaviors using departmental regulations and sanctions
1988	unknown	Program plagiarism revisited: current issues and approaches
1981	unknown	Plagiarism in computer sciences courses(Panel Discussion)
1973	unknown	On the structure of a computing profession
2009	Regular Article	A model academic ethics and integrity policy for computer science departments
2013	Regular Article	From difference to diversity: including women in the changing face of computing
2015	Abstract	Launching CROMA: Computational Research On Music & Audio
2016	Technical Note	How to Launch a STARS Computing Corps Cohort to Improve Retention and Broaden Participation in Computing (Abstract Only)
2017	Abstract	The Code of Ethics Quiz Show
2017	Abstract	The ACM Code of Ethics and Professional Conduct: Teaching Strategies and the Coming Update (Abstract Only)
2018	Regular Article	Key Concepts for a Data Science Ethics Curriculum
2018	Regular Article	Teaching Inclusive Thinking to Undergraduate Students in Computing Programs
2018	Abstract	Active Learning Strategies for Integrating the ACM Code of Ethics into CS Courses: (Abstract Only)
2018	Poster	Developing a Unique Android App-driven Nifty Middle-School Educational Module on Mobile Security for Driving Basic Information Security Awareness and Generating Interests in Cybersecurity: (Abstract Only)
2018	Poster	A Middle-School Code Camp Experience Emphasizing Data Science for Social Good: (Abstract Only)

Appendix B


Ethics Theme
Diversity
Discrimination
Ethic
Inclusive
Legal
2
1
0
-1
1970 1972 1974 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018 2020
Year

Appendix C


Academic Integrity Theme
Cheating
Conduct
Falsification
Integrity
Morals
Plagiarism
Wrongdoing
Year
1970
1972
1974
1976
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020