Brazilian School Girls' Perspectives on a Computer Science Major: Mining Association Rules

Maristela Holanda
Universidade de Brasília, Departamento de Ciência da Computação
Campus Darcy Ribeiro, Asa Norte, Brasília - DF, Brazil, 70910-900
mholanda@unb.br

and

Roberto N. Mourão
Universidade de Brasília, Departamento de Ciência da Computação
Campus Darcy Ribeiro, Asa Norte, Brasília - DF, Brazil, 70910-900
roberto.mourao@aluno.unb.br

and

Aleteia Araujo
Universidade de Brasília, Departamento de Ciência da Computação
Campus Darcy Ribeiro, Asa Norte, Brasília - DF, Brazil, 70910-900
aleteia@unb.br

and

Maria Emilia Walter
Universidade de Brasília, Departamento de Ciência da Computação
Campus Darcy Ribeiro, Asa Norte, Brasília - DF, Brazil, 70910-900
mariaemilia@unb.br

and

Vinicius R. P. Borges
Universidade de Brasília, Departamento de Ciência da Computação
Campus Darcy Ribeiro, Asa Norte, Brasília - DF, Brazil, 70910-900
viniciusrpb@unb.br

and

Guilherme N. Ramos
Universidade de Brasília, Departamento de Ciência da Computação
Campus Darcy Ribeiro, Asa Norte, Brasília - DF, Brazil, 70910-900
gnramos@unb.br

Abstract

The field of Computer Science has been of little interest for girls straight out of high school, when considering undergraduate majors in Brazil. At the University of Brasília's Department of Computer Science, female students compose less than 10% of the student body. In an effort to understand the girls' lack of interest in computer related courses, we applied an anonymous questionnaire, from 2011 to 2014, regarding their perceptions of the field. The participants were 3707 females students who completed an anonymous questionnaire. We applied Association Rules in Data Mining and Principal Component Analysis to analyze these responses . The knowledge gained through this study could guide future research on the matter and guidelines for motivating girls to pursue careers in Computer Science.

Keywords: computer science, gender, girls, women, data mining, association rules.

1 Introduction

In the years between 2000 and 2013, women accounted for an average of only 17% of all graduates in various Computer Science (CS) related majors in Brazil [1]. This situation is particularly distressing at one of Brazil's largest universities, the capital's University of Brasília (UnB), which has over 30,000 students enrolled in its undergraduate programs. There, over the last ten years, women have accounted for only 10% of all graduates in CS majors UnB [2].

Responding to these low rates in the number of women in CS courses, researchers have focused on how to improve this scenario, and have proposed strategies to encourage girls to pursue a profession in Computer Sciences [1, 2, 3, 4]. Several countries have developed initiatives to debate this issue, such as IEEE's *Women in Engineering*[1], ACM's *ACM-W*[2] and *Grace Hopper Celebration of Women in Computer Sciences*[3], the *Girls Who Code*[4] nonprofit organization, among others, all aiming to support and increase the number of women in CS.

The Brazilian Society of Computing has the *Meninas Digitais* project[5], originated from the *Women in Information and Technology* workshops held yearly since 2007 at the country's largest CS conference. The event was created to discuss this theme and Brazilian governmental agencies, such as the Ministry of Science and Technology, have been specifically calling for research projects related to the education of girls in STEM (science, technology, engineering and mathematics) majors [5].

Aiming to gather information on high school girls' perceptions of the field, UnB's CS Department started the `Meninas.comp` project[6], with the "Computer Science is girl's thing too" motto. This project's goals are to provide qualified information about the computer profession to high school students, to encourage the discussion about the lack of women in CS, to gather data about the process girls go through when choosing their profession, and to promote experimentation with computational activities.

One of the main challenges is understanding why girls do not want a career in CS in Brazil. Studies show that from the 1980s on, their interest in STEM careers decreased [6], and there is a perception of a male image of computer science in general [7]. In Brazil, there is a lack of research addressing the cause of so High School female students want the computing career so `Meninas.comp` elaborated a questionnaire of 14 questions about a career in CS to investigate. We applied it between the years of 2011 and 2014, polling 3707 girls, to examine their relationship with the computer as a tool and their interest in an undergraduate course in CS.

The resulting data was analyzed to examine the girls' affinity with the field, using the Apriori algorithm for searching for the association rules on their interest in CS and their background. Furthermore, an investigation of the relations among the most relevant and correlated variables describing the Brazilian girls is provided by applying the well-known Principal Component Analysis [8]. The goal is to understand these relationships and gain insights on the gender issue.

The rest of this paper is organized as follows: Section 2 presents related works, Section 3 provides details on the survey, Section 4 describes experimental results and findings, and concluding remarks are given in Section 5.

[1] http://wie.ieee.org

[2] http://women.acm.org/

[3] http://ghc.anitaborg.org

[4] https://girlswhocode.com

[5] http://meninas.sbc.org.br

[6] https://facebook.com/meninas.comp

2 Related Works

There have been several studies addressing the gender issue in Computer Science. Lagesen describes CS as STEM field that has excluded women [9]. Putnik et al. in [10] present data from Yugoslavia, comparing gender ratios and observing a higher number of men compared women.

Stout et al. in [11] and Cheryan et al. in [12] provide studies about stereotyping in Computer majors, also arguing that there is a higher ratio of men than women in this field in the US. Likewise, Mercier et al. present surveys, drawings, and interviews which were used to examine the perception of US middle school students about characteristics of knowledgeable computer users [7]. These results showed cultural stereotypes of a computer user: 89% were male and 94% wore glasses.

Keinan et al. show data that the ratio of graduated women from Bachelor's programs in CS was of almost 40% in 1984, dwindling to 20% by 2006 in the US [13]. Vardi has similar results, and adds that in 2013 and 2014, only 14.7% of those graduates were women [14].

Papastergiou used descriptive statistics, principal component analysis and analysis of variance in [15] to investigate 358 Greek high school students' intentions and motivation for pursuing academic studies in CS. This study looked into several factors, such as the influence of family and academic environment on their career choices, their perception of a professional career in CS, and their self-efficacy beliefs regarding computers. The analysis showed that a lack of exposure to and use of a computer at home and in school from early stages in the students' lives seems to be the main factor in discouraging them from studying CS, specially considering the data for girls.

Anderson et al. applied means, Mann-Whitney U test comparison, and non-parametric statistics in [16] to study possible factors related to low rates of female participation in education pathways leading to information and communications technology (ICT) professions, considering data from a three-year period. The survey had binary options, such as "I am very interested in computers" and "I am not interested in computers", was presented to 1,453 high school girls in their senior year. The study identified two factors associated with a woman's aversion to ICT careers: the perception that the subject is boring, and an intense dislike of computers.

Maia describes a similar situation in [1], presenting a study on female participation in university majors in CS in Brazil, based on the Higher Education Census data from the Ministry of Culture and Education between the years of 2000 and 2013. One of the issues raised was that while the number of male graduates increased 98% in the period, that of female graduates decreased 8%.

The Department of Computer Science[7] at the University of Brasília (UnB) currently offers three undergraduate degrees related to CS: Bachelor of Computer Science (since 1987), Licentiate in Computing (since 1997), and Computer Engineer (since 2008). Figure 1 shows gender data for students enrolled in them.

In 1987, when the Bachelor course began, the gender difference between students enrolled was relatively low: 47% of were female. However, this number decreased over time and, in 1997, this percentage fell to 10%; by 2013 it was only 6%. The Licentiate degree's ratio oscillates roughly around 11%, while the Computer Engineer, which already began with low numbers, saw them fall to less than 12% in the past three years.

We can see from the data that the difference between the number of male and female students enrolled in a CS major at UnB has increased over the years, similar to what was reported in surveys. This does not improve the current workplace outline, where there is a lower number of women in all computing occupations, which implies that our technology is being created by a relatively homogeneous group composed of male workers. Such pattern is especially troubling given ample evidence of the critical benefits diversity brings to innovation, problem-solving, and creativity [17].

Given this alarming decline, and the widening disparity between male and female representations in the field of Computer Science, our work aims to investigate possible reasons for such scenario in Brasilia, Brazil' Capital. With the knowledge acquired, we intend to propose actions to increase the entrance of girls in the CS related courses of our university.

3 Data Collection and Analysis

The enrollment of female students in Computer Science majors is decreasing every year and one of the biggest challenges in addressing this is to discover what motivates girls to avoid a CS major. Our research intends to further investigate the women's perceptions of the Computer Science field by looking at the prospective enrollees: girls in high school. There is a significant lack of research on this subject in Brazil, and we believe such analysis could aid in the proposal of policies for increasing female participation in the field.

[7] http://cic.unb.br

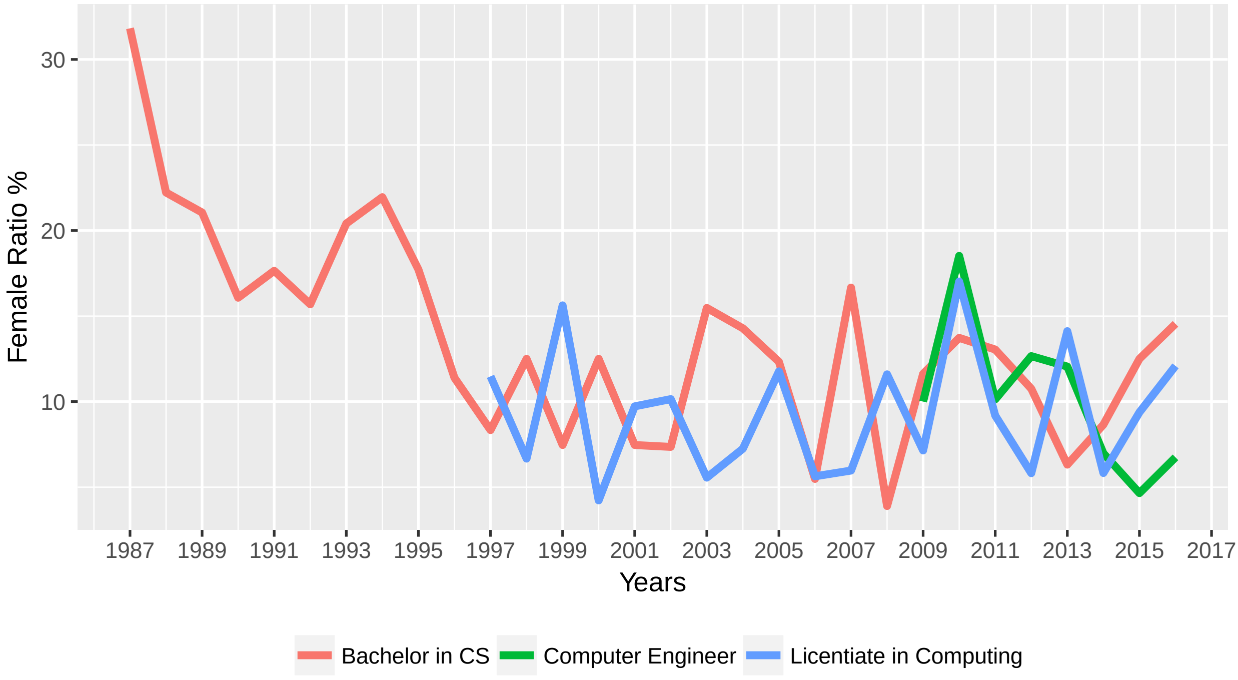


Figure 1: Ratio of female students enrolled in UnB's CS majors

3.1 Survey

The Meninas.comp Project from the University of Brasilia developed a questionnaire to inquire about female high school students' perceptions of the Computer Science field. It included personal profile questions, such as sex, school year, field of interest for a college education, interest in a career in CS, and others; as well as more general questions related to computers, such as: where the student uses computers and for what kind of tasks.

From 2011 to 2014, the questionnaires were given to female high school students in the Federal District, during Brazil's Ministry of Science and Technology's National Science and Technology Week[8] activities.

There were 1,821 responses in 2011, 944 in 2012, 517 in 2013, and 425 in 2014; adding up to a total of 3,707 completed questionnaires. The decrease of respondents in the period is due to the project's context, the amount of work possible is directly affected by the number of volunteers working on it. In 2011, there were 5 Professors and 10 students members, but they dwindles to only 1 Professor and 5 students (working few hours) in 2014. The collected data was consolidated in a spreadsheet and analyzed.

3.2 Data Analysis

Data analysis includes, among other things, procedures for analyzing data and techniques for interpreting their results [18], while Data Mining is the process of discovering insightful patterns and predictive models from data [19], in an effort to make sense of usually large amounts of information in some domain [20]. Our primary focus is the gender gap in STEM careers, so our study aims to characterize the profiles of girls who intend to enroll in undergraduate studies, especially those interested in Computer Science.

One of the possible approaches to finding interesting relationships in data is *association rules mining* [20], which produces easily understandable results as rules states as "*if A occurred, then B occurs*". For example, it is likely that a rule "*if a girls sees the CS field as boring, she will not enroll in a CS major*" is found. This processing may require a lot of resources, and a computationally feasible solution is the *Apriori* algorithm, which uses only the itemsets found large in the previous pass to generate candidate itemsets [21], and produces the rules with the highest *confidence* (how often the rule has been found to be true), despite their *support* (number of occurrences) [22]. The confidence is the conditional probability $P(B|A)$, i.e., the probability of A will occur, since B occurred [23]. In order to select which rules are more interesting rules, we consider their *lift*, which represents the level of association between the antecedent and the consequent [24].

An additional strategy for data analysis considers employing Principal Component Analysis (PCA) [8] to identify relations and patterns between the variables describing the Brazilian girls. Basically, PCA performs

[8] http://semanact.mcti.gov.br/

a linear transformation on data by combining high correlated attributes and obtaining uncorrelated new attributes. Such correlations can be obtained from the correlation (or covariance) matrix of the standardized Brazilian girls data. PCA works by decomposing the correlation matrix in order to obtain the eigenvalues and eigenvectors (principal components). The higher eigenvalues explaining the most of data variance are selected with their associated eigenvectors. Such eigenvectors defines the transformed representation of data and their interpretation can provide valuable information of the relations among attributes.

Thus, we search for insights on students' motivation for academic studies in Computer Science, with a clear gender bias (looking at only females) and analyzing a large data set of Brazilian students.

4 High School Girls' Perceptions On Computer Science

We took three approaches for analysis: statistical analysis for a better understanding of the data collected and association rules mining and PCA searching for interesting relationships between the data and the girl's interest in pursuing a CS major.

4.1 Statistical Analysis

The data for all years was consolidated in a single spreadsheet, which was then processed in the `R` programming language. The questionnaires, data and script used on this work are freely available online[9].

The preprocessing step cleans up the data (empty columns, whitespaces, etc.) and discards the data not in our subset of interest: Middle or High School girls who have answered whether they are interested in a CS major.

Figure 2 shows the respondents' interests in different scientific fields of undergraduate studies by year. The data indicates that, throughout the years, the percentages for each choice remains more or less the same, roughly around a value of 41% for *Biology-Health Sciences*, 22% for *Exact Sciences* and 33% for *Human Sciences*. The field related to Computer Science (Exact Sciences) is the least interesting for all years.

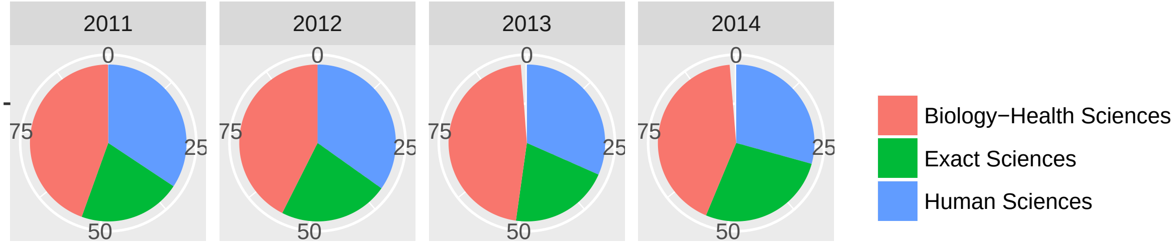


Figure 2: Respondents' interests in scientific fields

Figure 3 shows the respondents' interests in enrolling in a Computer Science major. On average, the data shows that 31% of the girls *have interest* while 28% *have no interest* and 41% *have doubt*. The data for the year 2011 differs a little from the others because the respondents for that year included Middle school students of all ages (from 5th to 9th graders). In the other years, only students from 9th grade or higher were surveyed due to limited human resources.

In order to investigate the profile of the students who are interested in Computer Sciences, we look at how the answers to the question *Would you major in Computer Science?* relate to the other questions. Figure 4 presents how students in different grades responded. The data shows that 12th graders had the lowest ratio of positive responses and that middle schoolers had the highest. This indicates an interesting research question: *why do girls lose interest in CS as they grow older?*

Figures 5 to 7 present the relationships of several variables observed. The titles indicate the question asked, and the legend on the right side the answers given; the plotted bars indicated how these answers relate do the respondents interest in enrolling in a CS Major.

Figure 5a and 5b show that the girls clearly know that CS majors teach more than just using softwares and requires Mathematical knowledge. Figure 5c shows that the majority of girls perceived that there are more boys than girls in CS majors. Within the group of girls that are not considering a CS major

[9] `http://goo.gl/oJYrjh`

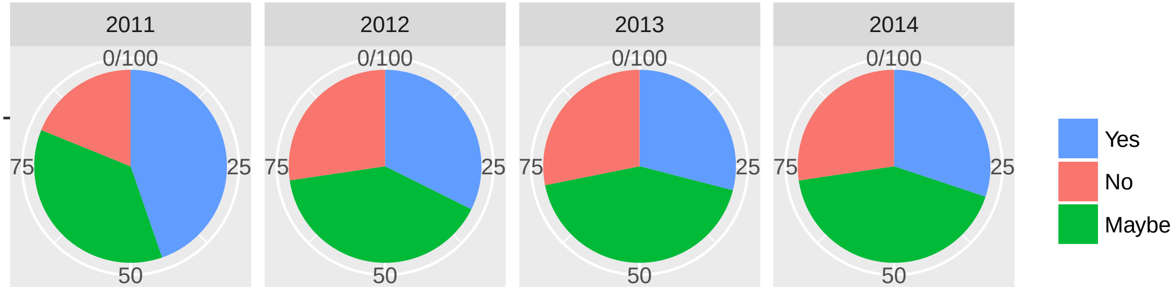


Figure 3: Respondents interested in Computer Science

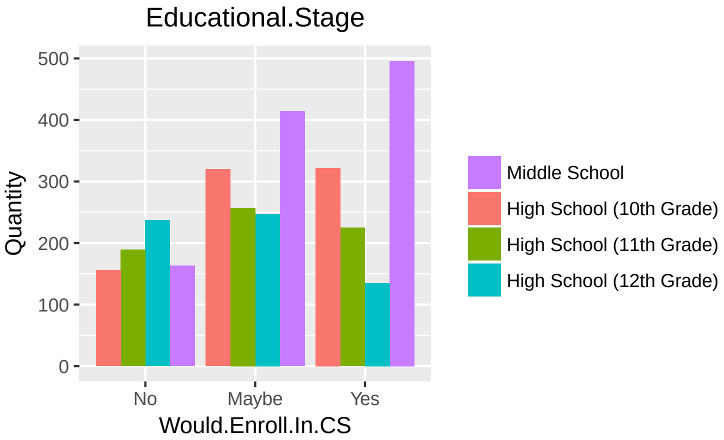


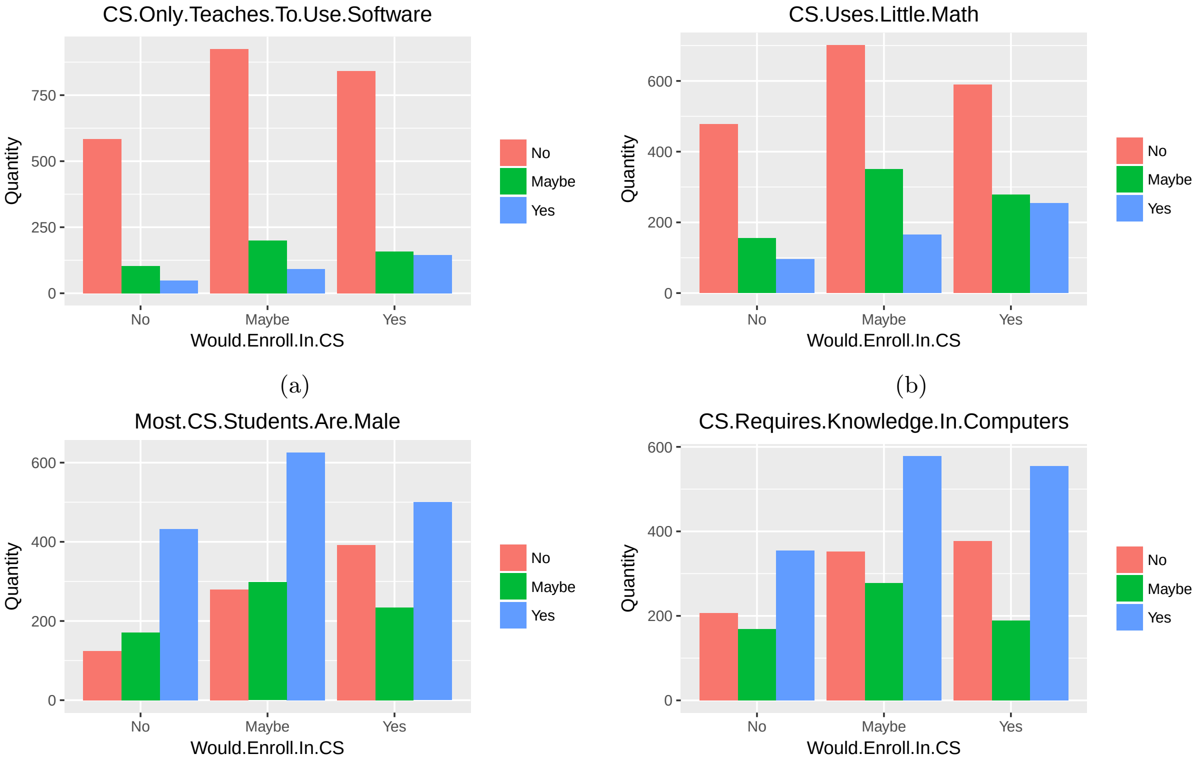
Figure 4: Results to the question *Would you major in Computer Sciences?* by Grade level

(`Would.Enroll.In.CS = no`), it is quite clear that they perceive the field to be dominated by men. This indicates an interesting research question: *are girls not interested in CS because most students are boys?* Figure 5d indicates that most of them perceive previous knowledge in using computers as a requirement for enrolling in a CS major, presenting another interesting question: *how much knowledge using computers is required to enroll?*

Figure 6a shows that the girls believe that a degree is required for a career in Computer Science. Figure 6b emphasizes the importance of family approval. Among the girls who answered that they would be interested in enrolling a major in CS, the majority reported having their family's approval (`Family.Approves.CS.Major = yes`); and the girls who reported not being interested had the highest rate of negative responses. Looking at Figures 6c and 6d, and considering the data for girls who did not say they wish to enroll, we wonder *why won't they pursue a career without long hours that they believe is full of opportunities?* Interestingly, the group of girls who reported being interested in CS, responded positively to *long hours* at a higher ratio than the other groups.

Figures 7a and 7d clearly show that the girls perceive CS as a creative field with various interdisciplinary possibilities. Figure 7b has a favorable perception; the group with the lowest *yes* response ratio to this question was that of girls who were not interested in CS, despite this group having a high ratio of *maybe* replies. Figure 7c indicates that the majority of girls think there are good salaries in the field, but it is worth to note that there was also a large part of them also responded *maybe*, specially in the group of girls who were not interested in CS.

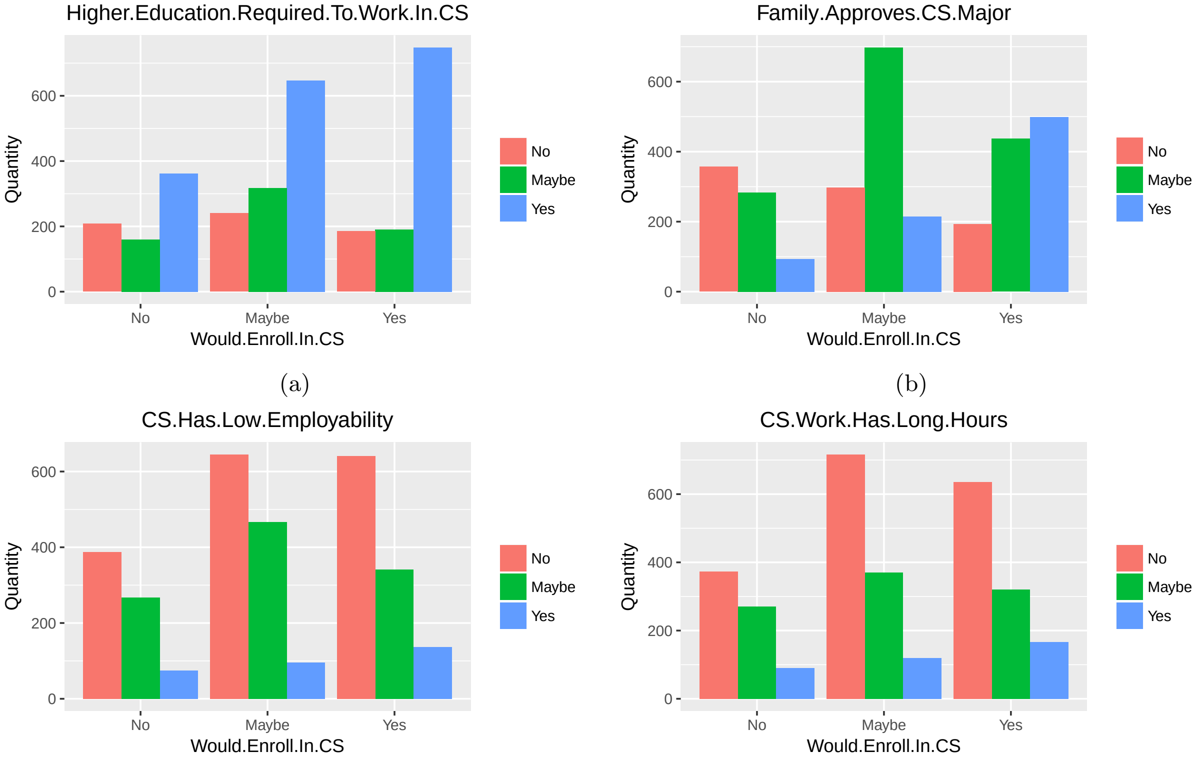
The remaining questions simply inquire where the girls use computers and what software tools they use. Almost all use a computer at home and most also use is at a relative or friend's house; about half use them at school, and the vast majority does not use a computer at work, at the library, or in digital inclusion centers. Considering tools, most have used text or image editors, but more than half have not used spreadsheets and very few have used databases.



(a) (b)

(c) (d)

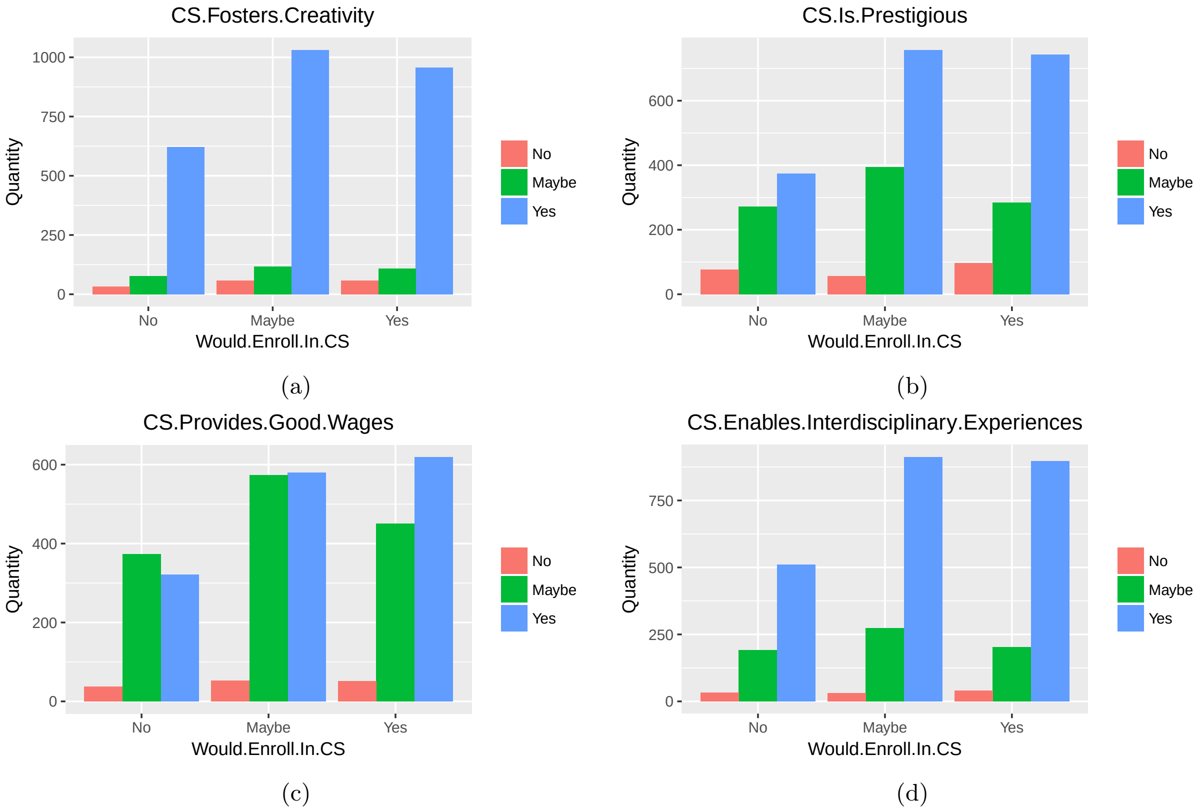
Figure 5: Relations between *Would.Enroll.In.CS* attributes and other variables



(a) (b)

(c) (d)

Figure 6: Relations between *Would.Enroll.In.CS* attributes and other variables



(a)

(b)

(c)

(d)

Figure 7: Relations between *Would.Enroll.In.CS* attributes and other variables

4.2 Mining Association Rules

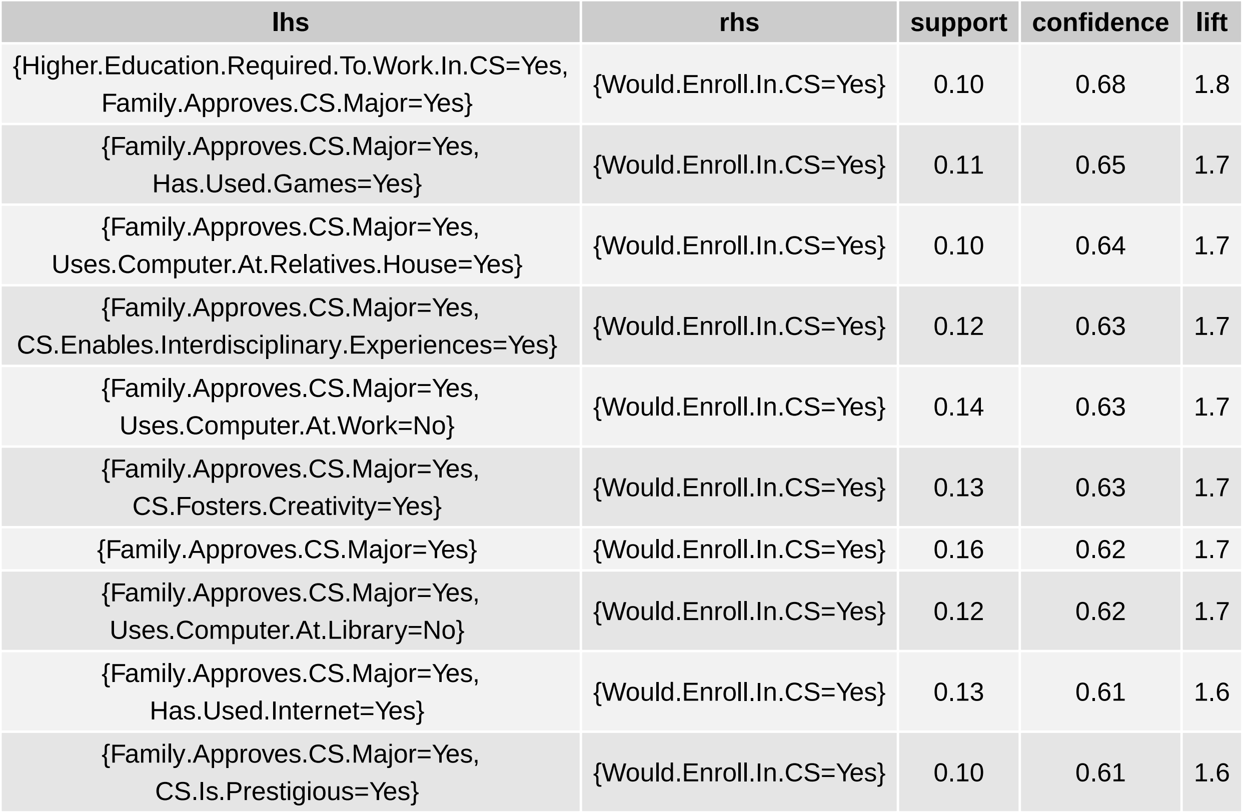
To gain further insights on the profiles of girls intending to enroll in undergraduate courses, especially those interested in ones related to Computer Science, we applied the Apriori algorithm to the data, with the minimum confidence level equal to 50%, and with a maximum number of 3 items in an itemset. This minimum confidence level removed rules with less than 50% of probability that the antecedent and the consequent were connected. The size of 3 items eased rules' interpretation, showing a maximum of two itemsets in the antecedent. A filter was applied to the rules' right-hand side in order to select only those involving the variable `Would.Enroll.In.CS`, which indicates the respondent is interested in pursuing a CS degree. The rules were analyzed considering *support*, *confidence* and *lift* metrics, and the association rule mining resulted in 32 rules.

We selected the 10 rules with highest lift for investigation, as detailed in Figure 8. They are consistent with our approach for the statistical analysis, all rules include the `Family.Approves.CS.Major` attribute. The rule with the highest lift also had the highest confidence level (of all rules); so we can say that there is 68% chance that "*the respondent would enroll in a CS major if she believes that she has her family's approval and that higher education is required to work with CS*". The lift value for this rule indicates that there is an 80% chance that its antecedents and the consequent are correlated, which is a very intriguing discovery: girls who are interested in and have their family's approval for a CS career *perceive the importance of their education.*

The rule with the second highest lift implies that, 65% of times, "*if the respondent believes that she has the family's approval and she has played computer games, then she would enroll in a CS major*". Again, there is a very strong correlation between the antecedents and the consequent indicating that girls who like computer games are more likely to pursue a CS Major.

This kind of analysis can be repeated for all the generated rules, providing several interesting insights on the profiles of girls who would enroll in a Computer Science course. These relationships can they be applied to guide actions addressing the gender issue in the field, as further discussed in Section 4.3.

Some rules that provided other discussion attention. The 5th rule, which reads as "*if the respondent believes that she has the family's approval and she does not use a computer at work, then she would enroll in a CS major*", has a strong correlation (70%) between the components, but an issue of ambiguity was raised on the attribute `Uses.Computer.At.Work=No`: what if the girls answered that they do not use a computer at work because they do not actually work? This makes it difficult to interpret the rule properly, since the



lhs	rhs	support	confidence	lift
{Higher.Education.Required.To.Work.In.CS=Yes, Family.Approves.CS.Major=Yes}	{Would.Enroll.In.CS=Yes}	0.10	0.68	1.8
{Family.Approves.CS.Major=Yes, Has.Used.Games=Yes}	{Would.Enroll.In.CS=Yes}	0.11	0.65	1.7
{Family.Approves.CS.Major=Yes, Uses.Computer.At.Relatives.House=Yes}	{Would.Enroll.In.CS=Yes}	0.10	0.64	1.7
{Family.Approves.CS.Major=Yes, CS.Enables.Interdisciplinary.Experiences=Yes}	{Would.Enroll.In.CS=Yes}	0.12	0.63	1.7
{Family.Approves.CS.Major=Yes, Uses.Computer.At.Work=No}	{Would.Enroll.In.CS=Yes}	0.14	0.63	1.7
{Family.Approves.CS.Major=Yes, CS.Fosters.Creativity=Yes}	{Would.Enroll.In.CS=Yes}	0.13	0.63	1.7
{Family.Approves.CS.Major=Yes}	{Would.Enroll.In.CS=Yes}	0.16	0.62	1.7
{Family.Approves.CS.Major=Yes, Uses.Computer.At.Library=No}	{Would.Enroll.In.CS=Yes}	0.12	0.62	1.7
{Family.Approves.CS.Major=Yes, Has.Used.Internet=Yes}	{Would.Enroll.In.CS=Yes}	0.13	0.61	1.6
{Family.Approves.CS.Major=Yes, CS.Is.Prestigious=Yes}	{Would.Enroll.In.CS=Yes}	0.10	0.61	1.6

Figure 8: Top 10 association rules for a girl's interest in a CS major, ordered by lift

straightforward understanding for this specific attribute is to check whether the contact with a computer at work, for girls who do work, has influence on her decision of enrolling in a CS major. This difficulty suggests improvements for our survey.

The 7th rule is also interesting: "*if the respondent believes that she has the family's approval then she would enroll in a CS major*" indicating (with a 70% confidence) that this attribute is important on it own. This rule's confidence and lift are only a fraction smaller than the previous ones, but its support is the largest of all rules.

We also provide an additional analysis of the relevant factors affecting the girls' decision to enroll in Computer Science. We applied PCA on the records concerning the Brazilian girls for 2014. This subset contains 425 instances, in which 72 of them are remove due to the presence of missing values. As PCA works on numeric data, a mapping procedure is performed in order to represent all categorical and nominal attributes to numerical ones. Table 1 exemplifies this procedure by considering a hypothetical attribute that can present three possible values, "Yes", "No" and "Maybe". In this procedure, such attribute values are mapped to the respective integer values, 0, 1 and 2.

Table 1: Mapping procedure example.

Attribute value	Integer value
"Yes"	0
"No"	1
"Maybe"	2

After preprocessing the original dataset, PCA is applied and the following outputs are produced: the eigenvalues and the eigenvectors (principal components). The number of selected eigenvectors were those covered by 95% of the total variance in the data according to their associated eigenvalues. Figure 9 illustrates the eigenvectors (each bar) and the associated eigenvalues (size of the bar), as well as the cumulative explained variance of data through the eigenvectors, depicted by the red line. It is clear that the first principal component accounts the higher amount of the total variation in the data, 12%. Moreover, there are several components associated to eigenvalues greater than one, which indicates that each one concentrates more information when compared to a unique variable. Those components are interpretable since some correlation between the components and original attributes can be found. Thus, there are several high

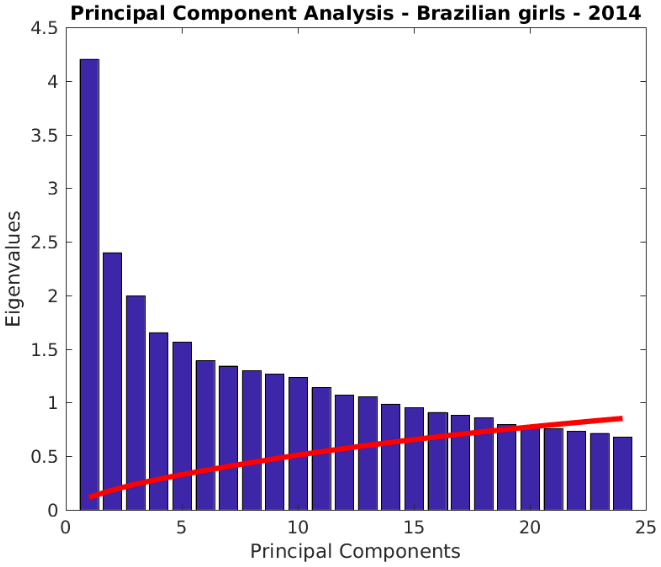


Figure 9: Principal component analysis: each bar is related to a principal component and their sizes reflect the eigenvalues. The red line indicates the cumulative explained variance.

correlated variables describing the Brazilian girls.

The strategy for interpreting the PCA's outputs is based on the determining the importance of principal components and the variables associated with each PC. Our analysis considers the two main principal components PC1 and PC2, which are associated to the higher eigenvalues. Each column in Table 2 describes the principal components 1 (PC1) and 2 (PC2), and the rows present the top five attributes that explains the higher correlations (in magnitude) for each principal component.

Table 2: Top-5 correlations for each attribute for the Brazilian girls in 2014.

PC1	PC2
Q2 (-0.41)	Uses.Computer.At.Lan.House (-0.328)
Q1 (-0.334)	Educational.Stage (0.309)
Has.Used.Database (-0.282)	Uses.Computer.At.Friends.House (-0.261)
Has.Used.Spreadsheet (-0.281)	Uses.Computer.At.Relatives.House (-0.251)
Has.Used.Image.Editor (-0.267)	Uses.Computer.At.Digital.Inclusion.Center (-0.25)

The analysis of PC1 on Table 2 shows that attributes Q1 and Q2 are relevant to that component, since they explain a considerable of the total variance of data when compared to the other data attributes. Moreover, the variables describing the use of computer softwares ("Has.Used.Database", "Has.Used.Spreadsheet", "Has.Used.Image.Editor") indicates a correlation between when explaining another amount of variance on PC1. The analysis of PC2 shows some correlation between the attributes describing girls using computers at different places, as reported in rows 1, 3, 4 and 5. However, the attribute "Educational.Stage" (row 2) also explains a significant amount of variance, but the positive signal expresses a contrast in relation to those attributes with negative signals.

4.3 Discussion of Results

The data analysis presented indicates that family approval is a very important factor in the choice of a major in higher education. We also saw that the majority of girls believe that there are more boys than girls majoring in CS, in accordance to stereotyping results in the United States [7], about stereotype in computing.

The employability in a CS career does not seem to be an important factor, though nearly 30% of the girls responded *maybe*, indicating that they are not sure about the job market in CS. This is a sensitive issue in Brazil's current context, the year 2016 ended with an unemployment rate of 12% [10] and informing the girls of their career possibilities might encourage them to pursue a CS major.

The need for spreading more information grows when all responses to the questionnaire are considered. The frequency of the answer *maybe* in the results was significant, as shown in Figure 3. This was specially

[10] http://www.ibge.gov.br/home/estatistica/indicadores/trabalhoerendimento/pnad_continua/default.shtm

true within answers from Middle School, as shown in Figure 4, hinting that motivating them earlier might lead to them becoming more interested in the area.

Another important point of notice is that the Brazilian girls are quite aware of the Math involved in a CS major. This notion, associated with the country's ongoing difficulties on teaching Math in schools, as reported by the National Institute for Educational Studies and Research [11], might be an important discouraging girls.

These findings and the discussions they sparked provide insights on factors that influence girls when choosing, or not, a major in Computer Science, which can direct efforts to mitigate the gender disparity. To this end, several possibilities are proposed. Universities can have activities disseminating information on the field, targeting the families of prospective, and possibly younger, students.

The field of CS can be brought into the students homes through their ubiquitous smartphones by developing software applications related to Math, Logic, and other CS notions to inspire the prospective students. Some of these applications could be games or incorporate gamification. CS could be closer to the girls realities through public policies, by creating more centers of digital inclusion in areas where girls have difficulties accessing a computer.

The analysis raised more interesting questions, which lead to more research inquiries and improvements on the questionnaire. It also showed how data mining can be useful in this kind of research, motivating us to apply other approaches to discover more knowledge.

5 Conclusion

In recent years, the gender gap in the field of Computer Science has widened and girls have not shown interest in majoring or becoming professionals. In order to address the issue, we performed a research to better understand this scenario and its causes, investigating the girls' perspectives on a Computer Science major. The knowledge discovered could to guide and support actions that will reduce the disparity by increasing participation of the girls.

Data mining and statistical analysis were applied to the data from a questionnaire done from 2011 to 2014, and results showed that the single most important factor for a Middle or High School girl deciding whether to pursue a degree in a CS major is her family's approval of this choice. Insights on the importance of other factors, such as career opportunities and Math requirements for a major were also obtained. The analysis and the mining process sparked new questions and discussions.

After this research, the project `Meninas.comp` changed some of its approaches. We created activities targeted exclusively at Middle School girls; provided lectures specifically about jobs in computing and lectures featuring important women in the field; proposed activities involving Mathematics, Computing and games. We will be watching the results on these actions to see if they will help reduce the gender gap.

Future works include: improving the questionnaire; applying the analysis to data for Middle and High School boys, for comparison, as well as to students who have enrolled in a CS major; and investigating other data mining approaches for knowledge discovery.

References

[1] M. M. Maia, "Limites de gênero e presença feminina nos cursos superiores brasileiros do campo da computação," *Cadernos Pagu*, vol. 0, no. 46, pp. 223–244, 2016.

[2] G. C. Couto and M. A. da Nóbrega Alberto Dantas, "Utilizando Mineração de Dados para Análise de gênero nos cursos de Computação na UnB," 2014. [Online]. Available: http://bdm.unb.br/handle/10483/8176

[3] J. M. Cohoon, "Recruiting and Retaining Women in Undergraduate Computing Majors," *SIGCSE Bull.*, vol. 34, no. 2, pp. 48–52, jun 2002.

[4] D. Gürer and T. Camp, "An acm-w literature review on women in computing," *SIGCSE Bull.*, vol. 34, no. 2, pp. 121–127, Jun. 2002.

[5] CNPq, "Chamada N° 18/2013 MCTI/CNPq/SPM-PR/Petrobras - Meninas e Jovens Fazendo Ciências Exatas, Engenharias e Computação." [Online]. Available: http://goo.gl/d8239T

[6] J. Abbate, *Recoding Gender: Women's Changing Participation in Computing*, 1st ed. One Rogers Street Cambridge MA 02142-1209: The MIT Press, 2012.

[11] `http://portal.inep.gov.br/web/guest/educacao-basica/saeb/resultados`

[7] E. M. Mercier, B. Barron, and K. M. O'Connor, "Images of self and others as computer users: the role of gender and experience," *Journal of Computer Assisted Learning*, vol. 22, no. 5, pp. 335–348, 2006.

[8] I. T. Jolliffe, "Principal component analysis and factor analysis," in *Principal component analysis*. Springer, 1986, pp. 115–128.

[9] V. A. Lagesen, "The strength of numbers: Strategies to include women into computer science," *Social Studies of Science*, vol. 37, no. 1, pp. 67–92, 2007.

[10] Z. Putnik, I. Štajner Papuga, M. Ivanović, Z. Budimac, and K. Zdravkova, "Gender related correlations of computer science students," *Computers in Human Behavior*, vol. 69, pp. 91 – 97, 2017.

[11] J. G. Stout, V. A. Grunberg, and T. A. Ito, "Gender roles and stereotypes about science careers help explain women and men's science pursuits," *Sex Roles*, vol. 75, no. 9, pp. 490–499, 2016.

[12] S. Cheryan, V. C. Plaut, C. Handron, and L. Hudson, "The stereotypical computer scientist: Gendered media representations as a barrier to inclusion for women," *Sex Roles*, vol. 69, no. 1, pp. 58–71, 2013.

[13] E. Keinan, "A New Frontier: But for Whom? An Analysis of the Micro-Computer and Women?s Declining Participation in Computer Science," Claremont McKenna College, Tech. Rep. 1466, 2017. [Online]. Available: http://scholarship.claremont.edu/cmc_theses/1466

[14] M. Y. Vardi, "What can be done about gender diversity in computing?: A lot!" *Commun. ACM*, vol. 58, no. 10, pp. 5–5, Sep. 2015.

[15] M. Papastergiou, "Are Computer Science and Information Technology still masculine fields? High school students' perceptions and career choices," *Computers & Education*, vol. 51, no. 2, pp. 594 – 608, 2008.

[16] N. Anderson, C. Lankshear, C. Timms, and L. Courtney, "'Because it's boring, irrelevant and I don't like computers': Why high school girls avoid professionally-oriented ICT subjects," *Computers & Education*, vol. 50, no. 4, pp. 1304 – 1318, 2008.

[17] C. Ashcraft, B. McLain, and E. Eger, *Women in tech: The facts.* Workforce NCWIT Alliance, 2016.

[18] J. W. Tukey, "The future of data analysis," *Ann. Math. Statist.*, vol. 33, no. 1, pp. 1–67, mar 1962.

[19] M. J. Zaki and J. Wagner Meira, *Data Mining and Analysis: Fundamental Concepts and Algorithms.* Cambridge University Press, May 2014.

[20] K. Cios, R. Swiniarski, W. Pedrycz, and L. Kurgan, "The Knowledge Discovery Process," in *Data Mining.* Springer US, Jan. 2007, pp. 9–24.

[21] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *VLDB94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, 1994, pp. 487–499. [Online]. Available: http://www.vldb.org/conf/1994/P487.PDF

[22] D. Taniar, W. Rahayu, V. Lee, and O. Daly, "Exception rules in association rule mining," *Applied Mathematics and Computation*, vol. 205, no. 2, pp. 735 – 750, 2008, special Issue on Advanced Intelligent Computing Theory and Methodology in Applied Mathematics and Computation.

[23] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics).* Greenwood Publishing Group, 2009.

[24] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, ser. Always learning. Pearson Addison Wesley, 2006.