Identifying Causal Structures from Cyberstalking: Behaviors Severity and Association

Shkurte Luma-Osmani, Florije Ismaili, Pankaj Pathak, and Xhemal Zenuni

***Abstract*—This paper presents an etiological cyberstalking study, meaning the use of various technologies and internet in general to harass or to stalk someone. The novelty of the paper is the multivariate empirical approach of cyberstalking victimization that has received less attention from the research community. Also, there is a lack of such studies from the causal perspective. It happens, since in most of the studies, a priority is given on a single causation identification, whereas the data examination used for mining causal relationships in this paper presents a novel and great potential to detect combined or multiple cause factors. The paper focuses in the impact that variables such as age, gender and the fact whether the participant has ever harassed someone, is related to the fact of being victim of cyberstalking. The research aims to find the causes of cyberstalking in high school's teenagers. Furthermore, an exploratory data analysis has been performed. A weak and moderate correlation between the factors on the dataset is emphasized. The odds ratio among the variables has been calculated, which implies that girls are twice as likely as boys to be cyberstalked. Similarly, concerning outcomes related to cyberstalking frequency recidivism are noticed.**

***Index Terms*—causality, causal rules, cyberstalking, data mining.**

I. Introduction

Causality is undoubtfully the center of any scientific approach and the experimental approach is best method for defining this occurrence. Anyway, considering several cost, law and ethical issues, observational data are considered a great replacement. In modern society everyone is connected to cyberspace in their daily life activities and even as an individual it's difficult to survive without being connected to technology. In the recent years more societal attention has been drawn towards cyberstalking victimization as a global matter. Along with the technology usage increasement, cybercrime is becoming the most aggressive, tech advanced, flourishing and fastest growing pattern of crime [18]. Adolescents, as the most targeted victims, lacking the agitation
about the nature and consequences of cybercrime, are being tangled in this kind of crimes in numerous ways [19].

Existing research has attempted to explore cyberstalking victimization along with identifying factors which increase the risk of being cyberstalked. There are many debates that are heard in the corridors about the causes of cyberstalking just like when developing new software [22].

First novelty of the manuscript relies on exploring the cyberstalking phenomenon from the causal point of view, since it has been very little or no explored at all by scientific researchers. Many cybercrimes occur, but not always the causes are known. Numerous authors list different causes that lead to such crime. However, the focal point of this study is the impact that age, gender and the fact if we have previously harassed someone, is a factor of being victim of cybercrime. The second novelty is the arrangement of the causes of cyberstalking in the form of mathematical equation.

The paper is organized as noted. The Introduction is followed by the related work in section II, and the paper novelty is presented in section III, whereas data analysis has been performed in the section IV. Correlation and the relationship among variables in the dataset are discussed in section V, whereas the main problem of causal discovery is presented in section VI. The paper highlights the limitations in section VII, followed by the conclusion.

II. Related Work

Thus, additional research is required to more explore cyberstalking victimization. Several authors [20] among the factors of different types of cybercrime, including cyberstalking, online child abuse and cyberbullying identified

Manuscript received August 24, 2021; revised November 21, 2021. Date of publication December 30, 2021. Date of current version December 30, 2021.

S. Luma-Osmani is a PhD Candidate on the Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, North Macedonia and on the same time she is engaged as a Research Assistant on the Faculty of Natural Sciences and Mathematics, University of Tetova, Tetovo, Republic of North Macedonia (e-mail: sl21455@seeu.edu.mk).

F. Ismaili and X. Zenuni are with the Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, North Macedonia (e-mails: f.ismaili@seeu.edu.mk, xh.zenuni@seeu.edu.mk).

Pankaj Pathak is with the Symbiosis Institute of Digital and Telecom Management, Symbiosis International (Deemed University), Pune, India (e-mail: pankajpathak@sidtm.edu.in).

Digital Object Identifier (DOI): 10.24138/jcomss-2021-0139

phenomena's such as: psychopathic behaviors, low self-control, social inequality, family income, less offline social life, unemployment etc. Into the bargain, Social Media Platforms facilitates [1] the communication with friends or to the general public. But on the other side, through these social media platforms the people can be targeted for malicious purposes and one of them is identity deception. The existing research has been carried out to identify identity deception detection. Behavioral evidence analysis may help in investigating digital [2] misconducts which involves human interaction between offender and victim. It also helps in better understanding the dynamics of specific misconduct.

The malicious content [3] can be analyzed based on classification of complex semantic events with ontology representation. An integrated approach [4] of social support from criminology and comorbidity to investigate correlation of stalking victimization has been conducted. The study concludes that individual life habits and social contexts both may responsible for being victimized. Cyberstalking is considered an anti-social problem and performed at a large scale [5] in various forms.

When considering the causality, several theoretical studies are developed aiming the exposure of cause-and-effect rules. Those published algorithms can be arranged into three sectors: proposing a novel approach algorithm [16][21], modifying the already existing algorithms or hybridizing numerous ones [17]. When answering the research question related to the most known causal rules mining methods, algorithms and techniques researchers in [15] list: LCD (Local Causal Discovery) algorithm along with its variants such as LCDa, LCDb and LCDc, PC (Peter and Clark), FCI (Fast Causal Inference), CCC, CCU, CAR (Causal Association Rule discovery), TC (Total Conditioning), CR-CS (Causal Rule mining with Cohort Study), CCCRUD (Conjunctive Combined Causal Rules Mining), DCCRUD (Disjunctive Combined Causal Rules Mining) etc.

III. Contribution of This Work

Cybercrime, as a growing phenomenon, undoubtedly represents a major field of study recently. Cyber stalking, as a part of it, can victimize all age groups, with a special emphasis on young people that can face sever forms of being harassed and monitored online. However, it's causes are not always known.

Those few works that have been focused on consequences, have dealt more with single causality factors, thus ignoring the fact that effects do not always come from a single cause, and there are cases when a variable cannot cause anything, but their combination can yield unexpected and very good results. After each iteration, a new variable is incorporated to the study and the odds ratio among the factors yield to concerning outcomes as respect to gender. Causal approach of cyberstalking victimization is noted, that can derive the essential elements for assessing it from different context among societies. Simultaneously, it is represented in a form of logistic model where three independent factors play an important role in causing a single output variable. Therefore, this paper presents discovering newly created rules based on observational data. On the other side, surveys, as a form of observational studies, often used to gather data from a sample group and consequently have an overview of the entire population. Different kind of surveys include distribution of questionnaires through mail, phone interviews, observations taken from house visits, censuses and similar. In all of the above-mentioned forms, the researcher has no impact on the result obtained. Therefore, a study can focus on factual thoughts or information, depending on the objective of the study [23]. Aiming a minor contribution on the open data society, the questionnaire results are publicly accessed on a Kaggle web repository.

IV. Exploratory Data Analyses

In North Macedonia, as in many other parts of the world, cybercrime is on the rise, and it is especially noticeable in the last decade. The dataset used in the paper is publicly published in the Kaggle database https://www.kaggle.com/shkurtelumaosmani /cyberstalking-in-tetovo [6]. Three high schools in Tetovo, Republic of North Macedonia have been surveyed and the data has been analyzed in Python programming language. The participants include 48.6% male and 51.4% female aged from 14 to 18 years old.

In the countplot below we have visualized forms of harassment for the people who have been victims of cyberstalking or who have not been victims. We can see that only two people were not victims and those two have reported that someone has ordered goods online with their data. More than 20 people who have been victimized said that the form of harassment was posting false information about them, this is also the most frequent form of harassment. The second most reported form of harassment (less than 20, more than 15 people), was classified as any other behavior found distressing in any way. The third form is threatened in chat rooms or comments, this counts for nearly 15 people. Threatening or abusive e-mails have reported around 5 people. There are less than 5 people who have been harassed with both forms, such as posting false information and also any other behavior found distressing in any way. Less than 3 have reported other combinations of other forms. In detail this could be observed by the countplot.

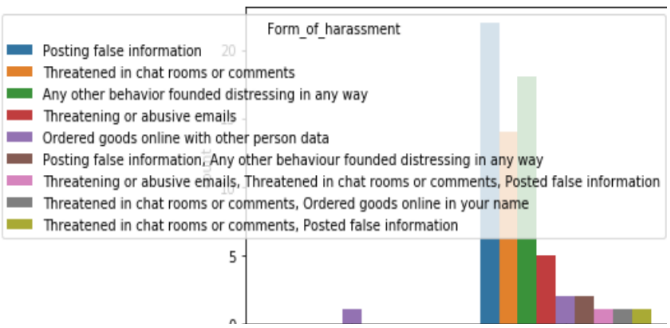

Fig. 1. Form of harassment

Intensity of cyberstalking in victims of cyberstalking varies. First let us explain the blue bar at the 0 side of the subplot. It means that one person that is not a victim of cyberstalking is harassed one time per month. This may be an error because the victims of cyberstalking have been requested this part while the

others no, or it may be that the person who answered was not felt as a victim of cyberstalking but anyways has been harassed at least once per month (or maybe just once in a lifetime). Higher number of people have been harassed 2 or 3 times per month (more than 12). Around 10 people have been harassed once per week. 9 people have been harassed every day and 7 people have been harassed every hour, which is very concerning.

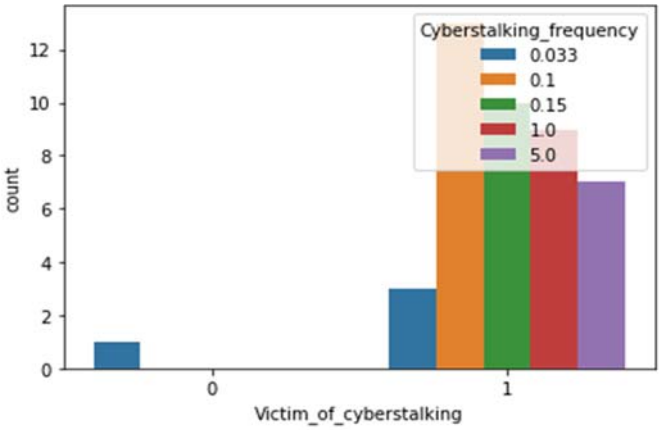

Fig. 2. Cyberstalking frequency

The boxplot addresses the cyberstalking frequency related to age of the surveyed people. As we know boxplots give us five summaries: the minimum, maximum, first quartile, median and the third quartile. The size of the box itself (which represents the interquartile range) shows that the given data have different dispersion around their median value. The 15 years old students have a box plot that is normal without outliers, the maximum frequency of cyberstalking in this group of age is 5 times and of course 0 is the minimum. The median is 2.5 because the number of observations is even (142 in total). The other 2 groups have outliers and all of the maximum and minimum values (times they have been cyberstalked) of both of them is 5 respectively 0.

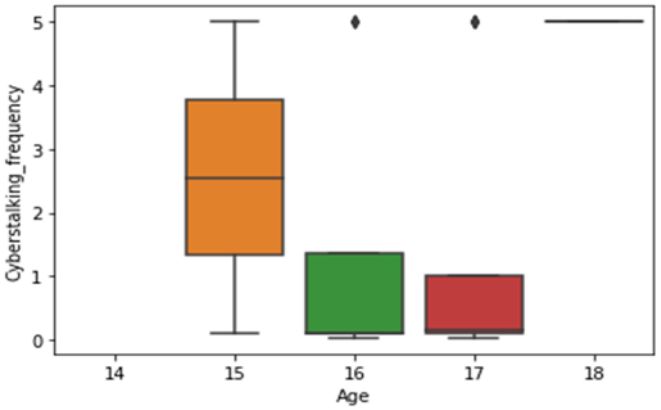

Fig. 3. Age based cyberstalking frequency

V. Correlation Analysis of Dataset Variables

In order to check the correlation between the variables in the dataset and the direction in which their linear relationship is (positive or negative), we can notice a significant correlation in a single plot named heatmap form the Seaborn library. However, correlations might or might not specify causal relationship [7][8]. Correlation shows us to what extend two variables are linearly associated. Linearity is seen when observing given coefficients of correlation. Down here is used the Pearson's correlation coefficient.

Python data visualizations through Matplotlib and Seaborn libraries presents a great way to quickly check the correlations between the columns in the dataset. Therefore, the stronger the shade of the color, the larger the magnitude of the correlation. Of course, the correlation of a column with itself results always in value 1, so called the perfect positive correlation. The values that appear closer to 0, mean that there is no linear trend detected among the two columns, whereas the closer to value 1 implies that variables are more positively correlated to each-other, it means that both of them will increase or decrease simultaneously, though the closer the coefficient value to -1 it has a meaning that variables are negatively correlated, i.e., one increases the other will decrease and vice versa.

Karl Pearson correlation coefficient [9] can indicate the level of the correlation among the variables, and can take the values, as per figure 4 below.

r value	Interpretation
+.70 or higher	Very strong positive relationship
+.40 to +.69	Strong positive relationship
+.30 to +.39	Moderate positive relationship
+.20 to +.29	weak positive relationship
+.01 to +.19	negligible relationship
0	No relationship [zero order correlation]
-.01 to -.19	negligible relationship
-.20 to -.29	weak negative relationship
-.30 to -.39	Moderate negative relationship
-.40 to -.69	Strong negative relationship
-.70 or higher	Very strong negative relationship

Fig. 4. Pearson correlation coefficient

The coefficient, indicating how strong this linear relationship is, can be calculated using the formula:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 - \sum(y_i - \bar{y})^2}}$$

where the variables represent as following: r - correlation coefficient, xi - values of the variable x in a sample, yi - values of the variable y in a sample, $\bar{x}$ - mean of the variable x, $\bar{y}$ - mean of the variable y.

We do have some significant correlations among the given variables. It would give a more detailed analysis the model built for dependency of variable of "Victim_of_cyberstalking". However, from the matrix we get the following:

The dependent variable "Victim_of_cyberstalking" helps us conclude that even though there are some signs of linear correlation to other independent variables, there are no strong coefficients of correlation (negative or positive). For instance, a moderate correlation between "Age" and "School" valued 0.36

is presented, there is also a slight positive linear correlation of 33% between victims of cyberstalking and getting rid of the cyber stalker. 27% is the positive linear correlation between the victims of cyberstalking and if you have harassed someone. Regarding the other variables, which are taken as independent to some linear analysis extent, there are noticeable coefficients between schools and if you harassed someone, 25% positive correlation.

It is interesting the negative correlation coefficient negatively valued -38% between variables "Victim_of_cyberstalking" and "Cyberstalking_achieving_goal". Gender and other variables have lower correlation coefficients to the victims of cyber stalking victims. "School" is also linearly associated to the "Cyberstalking_pleasure", the weak association is negative 26%. There are also two weak relationships between "Age" with, "Pleasure" and "School" with "Social_media_ communication", both equaled -0.23, meaning that the high school "Kiril Pejçinoviq" communicates more through social medias that "7 Marsi" and "Nikola Shtejn".

Higher coefficient of correlation, a positive of 47% have "Cyberstalking_pleasure" and "Cyberstalking_achieving_ goal". The result is expected since there is theoretically correlation between this two sociopathy measured variables, as displayed in detail on the heatmap 5 below.

All the coefficients are statistically significant for alfa=5% ($p<0.05$) except for the gender coefficient which is statistically significant for alfa=10% ($p<0.07$).

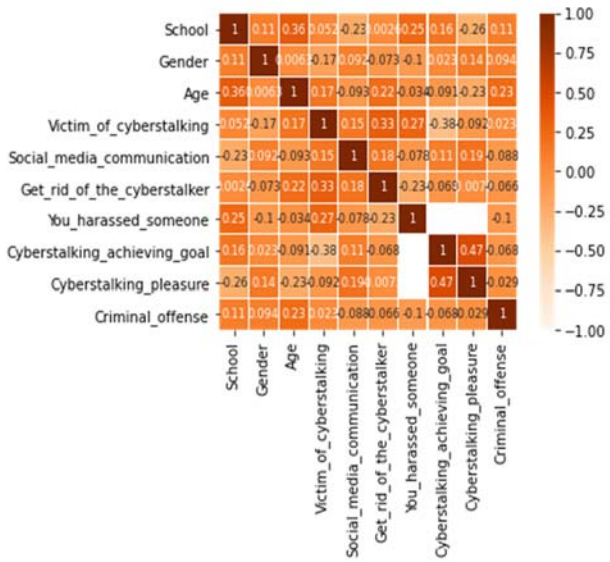

Fig. 5. Heatmap Correlation

Based on the results of the estimated correlation coefficient, no strong correlation among the defined variables has been noticed. This is because the positive correlation highest positive value is 0.47 and negative value is -0.26. Therefore, a weak and moderate relationship is emphasized.

VI. Discovering Causality

Bearing in mind that causal rules imply associations but the reverse does not hold anytime, our approach started by firstly generation of association rules, and after those causal rules were detected and analyzed. Apriori algorithm was used for generation of the association rules and the total number consisted of 446 association rules. The metric type was Confidence with 0.7 value, the lower bound for minimum support was set to 0.1, the upper bound for minimum support was 1.0 and delta factor value for iteratively decrease support was set to 0.05.

Consequently, the top 10 discovered association rules from the "Cyberstalking" dataset are listed in the Appendix of the paper.

Agreeing to this, many of the causal relationships that we're interested in, do not exhibit perfect relationships also man-made measurements are not perfect as a result scientific causal models are usually probabilistic in nature.

The concept of cause-and-effect must be operationalized as independent and dependent variables that can be measured, and it presents the first and foremost step in introducing the causal relationships in scientific research.

As per [10] the causal rule is presented as a combined cause of two or more binary variables, (X_1, X_2, …, X_n, Y), where the subsets of X present the causal variable and Y presents the effect. The characteristics of the methodology of using combined variables in finding causes, lies on the fact that those variables alone do not imply a causation, whereas their combination might.

Bayesian networks have been counted as a central contribution to the field of Artificial Intelligence (AI) in the last decade. Aiming to model the probability distribution of the conditional independence, the Bayesian networks uses the graphical demonstrations of the DAG's along with conditional probability tables (CPTs). In this depiction each line represents the conditional dependency i.e., the direct influence of one variable on another, and each node represents a distinctive random variable.

In order for the Bayesian network to model a probability it must satisfy the Markov Condition implicating that each variable is conditionally independent of its non-descendants, given its parents [7],[13]. Mathematically we can say:

$$P(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = P(X_{n+1} = x \mid X_n = x_n) \quad (1)$$

The hybrid approach presented in the paper, begins with the outcomes of the Apriori association rules. Therefore, there is an association between variables Gender and Victim_of_ cyberstalking, and the rule claims that: *Gender=Female → Victim_of_cyberstalking=Yes.*

TABLE I. Two Variables Ratio

	Victim of Cyberstalking = Yes	Victim of Cyberstalking = No
Gender = F	40/73 = 0.548	33/73 = 0.452
Gender = M	26/69 = 0.377	43/69 = 0.623

The ratio between being female and victim of cyberstalking towards not being victim of cyberstalking is 1.21:1, whereas in the male group this ratio is 0.60:1. Namely, the odds for a female being victim of cyberstalking is 1.21 and the odds for a male being victim of cyberstalking is 0.60. It means that females have a double probability of being cyberstalked as compared to males. Of course, when the value of odd ratio is 1, it means that a variable has an equal probability to appear in both gender groups which does not appear in our case.

In the countplot bellow we can observe closely the count of victims of cyberstalking by gender. The discrepancy in the length of the bar is obvious and it has a mirrored form in some sort of way. Those people that have not been cyberstalked are more males than females and vice versa, those victims of cyberstalking are more females than males. The generated rule is also graphically represented in figure 6.

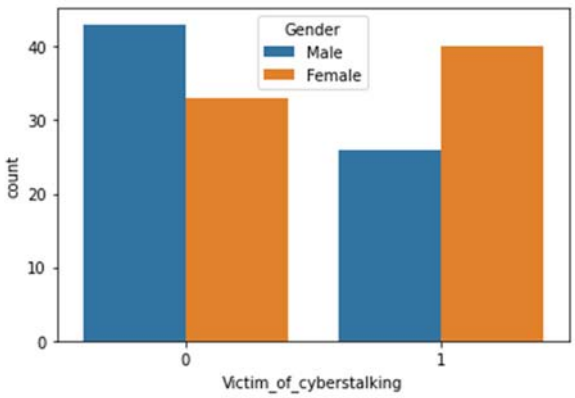

Fig. 6. Victim of Cyberstalking variable

If we go deeper into the analysis and add another attribute, You_harassed_someone, then we gain the resulting table:

TABLE II. THREE VARIABLES RATIO

	Victim of Cyberstalking = Yes	Victim of Cyberstalking = No
Gender = F & You_harassed_someone = Yes	23/34=0.676	11/34=0.324
Gender = F & You_harassed_someone = No	17/39=0.436	22/39=0.564
Gender = M & You_harassed_someone = Yes	14/25=0.56	11/25=0.44
Gender = M & You_harassed_someone = No	12/44=0.273	32/44=0.727

Therefore, if we relate to the third variable "Victim_of_cyberstalking" we notice that 68% of women who have harassed someone, have been victim of such phenomena, versus 32% who have not been a victim. On the contrary, 44% of females who have never harassed someone have been victim of cyberstalking and 56% of them have never neither harassed someone or being a victim.

As related to males, 56% of them have already harassed someone, and at once been a victim of cyberstalking, and 44% have done such harassment, but not been a victim. And finally, 27% haven't harassed anyone, but anyway experienced a stalking and 73% haven't done harassment or had such a consequence.

Recently, let us add one more variable "Age" to see how the same influences in the variable "Victim_of_cyberstalking". Relevant rules show that with age, the chance of being more victimized also increases.

[11] states that the difference between the expected value of conditional Y on causal rules σ if they are true or not, in the following formula where the empirical estimator *e* is calculated:

$$e\,(\sigma) = E[\,\mathrm{Y} \mid \sigma = \mathrm{T}] - E[\,\mathrm{Y} \mid \sigma = \mathrm{F}] \quad (2)$$

TABLE III. FOUR VARIABLES RATIO

Victim of Cyberstalking	Yes	No
Gender = F, Age = 14 & You_harassed_someone = Yes	0	0
Gender = F, Age = 15 & You_harassed_someone = Yes	1/5=0.2	4/5=0.8
Gender = F, Age = 16 & You_harassed_someone = Yes	3/3=1	0
Gender = F, Age = 17 & You_harassed_someone = Yes	19/26=0.73	7/26=0.27
Gender = F, Age = 18 & You_harassed_someone = Yes	0	0
Gender = F, Age = 14 & You_harassed_someone = No	0	0
Gender = F, Age = 15 & You_harassed_someone = No	1/5=0.2	4/5=0.8
Gender = F, Age = 16 & You_harassed_someone = No	3/4=0.75	1/4=0.25
Gender = F, Age = 17 & You_harassed_someone = No	11/27=0.40	16/27=0.593
Gender = F, Age = 18 & You_harassed_someone = No	2/3=0.667	1/3=0.333
Gender = M, Age = 14 & You_harassed_someone = Yes	0	0
Gender = M, Age = 15 & You_harassed_someone = Yes	0	3/3=1
Gender = M, Age = 16 & You_harassed_someone = Yes	2/2=1	0
Gender = M, Age = 17 & You_harassed_someone = Yes	12/20=0.6	8/20=0.4
Gender = M, Age = 18 & You_harassed_someone = Yes	0	0
Gender = M, Age = 14 & You_harassed_someone = No	0	1/1=1
Gender =M, Age = 15 & You_harassed_someone = No	0	5/5=1
Gender = M, Age = 16 & You_harassed_someone = No	3/5=0.6	2/5=0.4
Gender = M, Age = 17 & You_harassed_someone = No	8/29=0.276	21/29=0.724
Gender = M, Age = 18 & You_harassed_someone = No	1/4=0.25	3/4=0.75

There are a lot of authors that have made more known and identified the concepts that have to do with causality by using probability distributions defined on directed acyclic graphs [12] [13] [14]. We are more concentrated on the explanation given by [11] because it has also taken into consideration the explanation of the other authors as well.

The triangle done for the causal analysis in this research is built by using the following rules and connections:

- Y is the binary target variable and, in our circumstance, it is our victims of cyberstalking variable.
- X_1, X_2 and X_3 as a subset of X are description variables and, in our case, we have three of them: gender, age and the fact if you have harassed someone or not.

So, we can write the following based on the prior information: Y = {0,1}, and we know that it mathematically represents the domain of Y. On the other hand, the domain of X_i is either real or categorical. In our case it is both. [11] claim that the domain of X is an M dimensional outer product space $X= X_1 \times X_2 \times \dots \times X_m$.

For the sigma causal rules mentioned in the equation (1) we can define "an unbiased intervention do(σ) as the randomized operation of satisfying σ by setting X_σ to some x such that σ(x)=True according to the probabilities p (X_σ=x | σ=True)" and one needs to find rules of causal reasoning sigma that maximize the causal effect defined as the difference of expected value of Y which would be under two "interventions" do(σ) and do(¬ σ):

$$e(do\ (\sigma)) = E\ [Y \mid\ do\ (\sigma)] - E\ [Y \mid do\ (\neg\ \sigma)] = p(Y = 1 \mid\ do\ (\sigma)\ ) - p\ (Y = 1 \mid do\ (\neg\ \sigma))\ \ (3)$$

The method used in proceed is based on Bayesian network as a technique that marked a giant leap in causal discovery [15] where each additional node on the problem is based on the conditional probability distribution table as presented on the below network, where the variables are:

- X_1 – Gender (0: Female, 1: Male)
- X_2 - Knowing whether the participant harassed someone (0: No, 1: Yes)
- X_3 - Age (0: Doesn't have 17 years, 1: Has 17 years)
- Y - Victim of cyberstalking (0: No, 1: Yes)

Taking in consideration the aforementioned variables, an empirical logit model in aforementioned variables, as described in [6] is represented as follows:

$$f\ (x_1, x_2, x_3) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$
$$\widehat{Victim\ of\ Cybst} = -9.534 - 0.6598\ (Gender) + 0.552\ (Age) + 1.177\ (You\ \_harassed\_someone) \quad (4)$$

The standard error of each coefficient separately includes: Gender 0.3536, Age 0.2500 and the variable You harassed someone 0.3688, whereas the 95% level of confidence interval is as noted below:

Gender [-1.3695; 0.0500]
Age [0.0615; 1.0416]
You _harassed_someone [0.4546; 1.9003]

Spirtes et al. in the book [13] emphasize that studies related to various experiments and observations do not lead always to the same inferences and conclusions. Moreover, these consents show that there is a strong connection between causality and probability and this connection can help many more topics in statistic as a comparative power of the observation despite experiment, or as the Simpson's paradox is known, errors in regression models, sampling and variable selection.

VII. Limitations of The Study

It should be borne in mind that the obtained survey results will not provide information for the overall attitude of the adolescents, since the research was conducted only with teenagers in three high schools in the municipality of Tetova. The sample is limited and even though it is randomly chosen there should be clear the fact that the results presented here are only for adolescents (limited age range), Albanian and in the region of Tetovo (region centered). The results should not be biased if we take into consideration the abovementioned facts.

TABLE IV. CYBERSTALKING BAYESIAN NETWORK

$p(X_1=0)$	$p(X_1=1)$
0.51	0.49

X_1

X_1	$p(Y=1\|X_1)$	$p(Y=0\|X_1)$
0	0.55	0.45
1	0.38	0.62

X_2

X_1	X_2	p (Y=1 | X_1, X_2)	p (Y=0 | X_1, X_2)
0	1	0.68	0.32
0	0	0.44	0.56
1	1	0.56	0.44
1	0	0.27	0.73

X_3

Y

X_1	X_2	X3	p (Y=1 | X_1, X_2, X_3)	p (Y=0 | X_1, X_2, X_3)
0	1	0	0	0
0	1	0	0.2	0.8
0	1	0	1	0
0	1	1	0.73	0.27
0	1	1	0	0
0	0	0	0	0
0	0	0	0.2	0.8
0	0	0	0.75	0.25
0	0	1	0.41	0.59
0	0	1	0.67	0.33
1	1	0	0	0
1	1	0	0	1
1	1	0	1	0
1	1	1	0.6	0.4
1	1	1	0	0
1	0	0	0	1
1	0	0	0	1
1	0	0	0.6	0.4
1	0	1	0.28	0.72
1	0	1	0.25	0.75

VIII. Conclusion

No strong correlation coefficients among variables have been noticed. Several studies claim that there is a strong link between probability and causality. Association rules were firstly generated through Apriori algorithm and a ratio between three independent variables: "Gender", "Age",

"You_harassed_someone" and the dependent variable "Victim of Cyberstalking" has been calculated. The model analyzes data on how various factors have caused the behavior of Cyberstalkers. If you are a female according to this scheme the chances are 54.8% approximately to be harassed online. Boys are more likely not to be harassed, even with 62.3% approximately, i.e., the probability of being bullied and being boy is 37.7% respectively. Lastly, the likelihood of being cyberstalked increases, if you have previously harassed someone. The same consequence comes with increasing the age of the participants.

APPENDIX A
ASSOCIATION RULES

1. Gender=Female Get_rid_of_the_cyberstalker=Yes 36 ==> Victim_of_cyberstalking=Yes 36 <conf:(1)> lift:(2.15) lev:(0.14) [19] conv:(19.27)
2. Social_media_communication=Yes 41 ==> Victim_of_cyberstalking=Yes 41 <conf:(1)> lift:(2.15) lev:(0.15) [21] conv:(21.94)
3. Get_rid_of_the_cyberstalker=Yes 59 ==> Victim_of_cyberstalking=Yes 59 <conf:(1)> lift:(2.15) lev:(0.22) [31] conv:(31.58)
4. Social_media_communication=Yes Get_rid_of_the_cyberstalker=Yes 38 ==> Victim_of_cyberstalking=Yes 38 <conf:(1)> lift:(2.15) lev:(0.14) [20] conv:(20.34)
5. Get_rid_of_the_cyberstalker=Yes Criminal_offense=Yes 38 ==> Victim_of_cyberstalking=Yes 38 <conf:(1)> lift:(2.15) lev:(0.14) [20] conv:(20.34)
6. Cyberstalking_achieving_goal=Yes 37 ==> You_harassed_someone=Yes 37 <conf:(1)> lift:(2.41) lev:(0.15) [21] conv:(21.63)
7. Cyberstalking_pleasure=No 33 ==> You_harassed_someone=Yes 33 <conf:(1)> lift:(2.41) lev:(0.14) [19] conv:(19.29)
8. Social_media_cyberstalking=Instagram 31 ==> Victim_of_cyberstalking=Yes 31 <conf:(1)> lift:(2.15) lev:(0.12) [16] conv:(16.59)
9. Get_rid_of_the_cyberstalker=Yes You_harassed_someone=Yes 31 ==> Victim_of_cyberstalking=Yes 31 <conf:(1)> lift:(2.15) lev:(0.12) [16] conv:(16.59)
10. Social_media_cyberstalking=Instagram Get_rid_of_the_cyberstalker=Yes 30 ==> Victim_of_cyberstalking=Yes 30 <conf:(1)> lift:(2.15) lev:(0.11) [16] conv:(16.06)

REFERENCES

[1] Van der Walt, E., & Eloff, J. Identity deception detection: requirements and a model. *Information and Computer Security*, 26(4), 562–574, 2019. https://doi.org/10.1108/ICS-01-2019-0017

[2] Al Mutawa, N., Bryce, J., Franqueira, V. N. L., Marrington, A., & Read, J. C. Behavioural Digital Forensics Model: Embedding Behavioural Evidence Analysis into the Investigation of Digital Crimes. *Digital Investigation,* 28, 70–82, 2019.https://doi.org/10.1016/j.diin.2018.12.003

[3] Sobhani, F., & Straccia, U. Towards a forensic event ontology to assist video surveillance-based vandalism detection. CEUR Workshop Proceedings, 2396, 30–47, 2019.

[4] Cao, L., & Kevin Wang, S. Y. Correlates of stalking victimization in Canada: A model of social support and comorbidity. *International Journal of Law, Crime and Justice*, 63(September), 100437, 2020. https://doi.org/10.1016/j.ijlcj.2020.100437

[5] Abu-Ulbeh, W., Altalhi, M., Abualigah, L., Almazroi, A. A., Sumari, P., & Gandomi, A. H. Cyberstalking victimization model using criminological theory: A systematic literature review, taxonomies, applications, tools, and validations. *Electronics* (Switzerland), 10(14), 2021. https://doi.org/10.3390/electronics10141670

[6] Luma-Osmani, S., Ismaili, F., Ram Pal, P., "Building a model in discovering multivariate causal rules for exploratory analyses", IEEE International Conference on Data Analytics for Business and Industry (ICDABI), Bahrain. pp. 272-276, 2021. DOI: 10.1109/ ICDABI53623.2021.9655981.

[7] Cooper, G. F. A Simple Constraint-Based Algorithm for Efficiently Mining Observational Databases for Causal Relationship. *Data Mining and Knowledge Discovery*, 1(2), 203–224, 1997. https:/doi.org/ 10.1023/ A:1009787925236

[8] Wang, J., & Mueller, K. The Visual Causality Analyst: An Interactive Interface for Causal Reasoning. IEEE Transactions on Visualization and Computer Graphics, 22(1), 230-239, 2016. https://doi.org/10.1109/ TVCG.2015.2467931

[9] ElegantJ BI. (Smarten, Producer) Retrieved March 30, 2021, from https://www.elegantjbi.com/blog/what-is-karl-pearson-correlation-analysis-and-how-can-it-be-used-for-enterprise-analysis-needs.htm, 2018.

[10] Li, J., Le, T. D., Liu, L., Liu, J., Jinyz, Z., & Sun, B. Mining Causal Association Rules. IEEE 13th International Conference on Data Mining Workshops 114–123, Texas, 2013. https://doi.org/10.1109/ICDMW. 2013.88

[11] Budhathoki, K., Boley, M., & Vreeken, J. Rule Discovery for Exploratory Causal Reasoning. 32nd Conference on Neural Information Processing Systems (NIPS2018). Canada, 2018.

[12] Shimizu, S., Hoyer, P. O., Hyvärinen, A., & J., K. A. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research,* 7, 2003-2030, 2006.

[13] Spirtes, P., Glymour, C., & Scheines, R. Causation, Prediction and Search 2nd Edition. Massachusetts, USA, 2000.

[14] Pearl, J. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2), 2010. https://doi.org/10.2202/1557-4679.1203

[15] Luma-Osmani, S., Ismaili, F., Zenuni, X., & Raufi, B. A Systematic Literature Review in Causal Association Rules Mining. In P. R. Paul (Ed.), 11th Annual IEEE Information Technology, Electronics and Mobile Communication Conference-IEMCON, 20-26, Vancouver, BC, Canada, 2020. https://doi.org/10.1109/IEMCON51383.2020.9284908.

[16] Abualigah, L., Yousri, D., Abd Elaziz, M., Ewees, Ahmed A., Al-Qaness, Mohammed A.A., Gandomi, Amir. H. Aquila Optimizer: A novel meta-heuristic optimization algorithm, *Computers & Industrial Engineering*, 157, 2021. doi.org/10.1016/j.cie.2021.107250

[17] Abualigah, L., Diabat, A., Mirjalili, S., Abd Elaziz, M., Gandomi, Amir. H. The Arithmetic Optimization Algorithm, Computer Methods in Applied Mechanics and Engineering, 376, 2021. doi.org/10.1016/j.cma. 2020.113609

[18] Meško, G. On Some Aspects of Cybercrime and Cybervictimization. *European Journal of Crime, Criminal Law and Criminal Justice*, 26(3), 189–199, 2018.

[19] Shabnam, N., Faruk, M. O. and Kamruzzaman, M. Underlying Causes of Cyber-Criminality and Victimization: An Empirical Study on Students. *Social Sciences*. Vol. 5, No. 1, pp. 1-6, 2016.

[20] Abdullah, A. & Jahan, I. Causes of Cybercrime Victimization: A Systematic Literature Review. *International Journal of Research and Review*. 7(5), pp.89- 98, 2020.

[21] Ram Pal, P., Pathak, P., & Luma-Osmani, S. IHAC: Incorporating Heuristics for Efficient Rule Generation & Rule Selection in Associative Classification. *Journal of Information & Knowledge Management*, 20 (01), 2150010 - 1-13, 2021. doi:10.1142/S0219649221500106.

[22] Luma-Osmani, S., Arifi, G., Idrizi, F. Choosing the Most Suitable Model for Developing a Software. Sixth International Conference on Computational Intelligence, Communication Systems and Networks. pp. 83-88, 2014. doi: 10.1109/CICSyN.2014.30

[23] Roberts, D., Collecting Data: Surveys, Experiments, & Observational Studies. (n.d.). (Retrieved March 10, 2021, from https://mathbits notebook.com/Algebra2/Statistics/STSurveys.html.

Shkurte Luma-Osmani is presently pursuing the Ph.D. degree on the Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, North Macedonia. From March 2014 is engaged in the Office for Scientific Research and Innovation of the University of Tetova, whereas from September 2014, also works as a Research Assistant on the department of Informatics, Faculty of Natural Sciences and Mathematics, University of Tetova, North Macedonia. She is a member of International Association of Educators and Researchers (IAER), Kemp House, London, EC1V 2NX, UK.

Florije Ismaili works as an Associate Professor in the Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, North Macedonia. She was awarded with the title Doctor of Computer Sciences in the Technical University of Sofia, in the city of Sofia, Bulgaria in 2011 with thesis "Enhanced Web Service Discovery by using Data Mining Techniques, Latent Semantic Indexing and Semantic Web". Currently she's teaching Distributed Systems, Data Mining, Web Development and Web Services.

Pankaj Pathak is currently working as an Assistant Professor, in the Symbiosis Institute of Digital and Telecom Management, Symbiosis International (Deemed University) in Pune, India. His research scientific field include Data Mining and Decision Making. Simultaneously, he is an author and coauthor of several Research papers that are published in National and International level journals.

Xhemal Zenuni works as an Associate Professor in the Faculty of Contemporary Sciences and Technologies, South East European University, Tetovo, North Macedonia. He finished his Ph. D degree in the French Language Faculty of Electrical Engineering, in the Technical University of Sofia, in Sofia, Bulgaria. His thesis was entitled "QoS Aware Semantic Service Composition via AND/OR Graphs".