RAYUELA
a fun way to fight cybercrime

Deliverable Report

D6.4 Open Report: Profile Matching and Risk Indicators For Potential Young Victims

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 882828. The sole responsibility for the content of this document lies with the author and in no way reflects the views of the European Union.

Deliverable No.	D6.4	Work Package No.	WP6	Task/s No.	Task 6.3
Work Package Title		DATA ANALYSIS AND INTERPRETATION ON PROFILES FROM POTENTIAL YOUNG VICTIMS AND OFFENDERS			
Linked Task/s Title		T6.3 Profile matching and definition of risk indicators for potential young victims and offenders			
Status		Final	(Draft/Draft Final/Final)		
Dissemination level		PU	(PU-Public, PP, RE-Restricted, CO-Confidential)		
Due date deliverable		30/09/2023	Submission date		30/09/2023
Deliverable version		1.0			

Document Information and contributors

Deliverable responsible		COMILLAS	
Contributors	Organisation	Reviewers	Organisation
Jaime Pérez	COMILLAS	Kaisa Kägu	EPBG
Gabriel Valverde	COMILLAS	Catherine Monbailliu	UGENT
Mario Castro	COMILLAS	Violeta Vázquez	ZABALA
Gregorio López	COMILLAS	Abel Muñiz	ZABALA
Lokesh Sharma	NEC		
Sonia Solera	UPM		
Manuel Álvarez-Campana	UPM		

Document History

Version	Date	Comment
0.1	24/07/2023	First draft – Definition of structure and content
0.2	31/07/2023	First complete draft
0.3	31/08/2023	Improved draft ready for review
1.0	15/09/2023	Final version including reviewers’ comments

Table of contents

List of Abbreviations

Abbreviation	Description
BN	Bayesian Network
CB	Cyberbullying
CH	Cyber Harassment
DAG	Direct Acyclic Graph
OG	Online Grooming
t-SNE	t-Distributed Stochastic Neighbour Embedding
WP	Work Package

Executive Summary

This deliverable represents the last outcome of Task 6.3 "Profile matching and definition of risk indicators for potential young victims and offenders", which is the third task of Work Package (WP) 6 in the RAYUELA project. This task obtains as input (from previous tasks) a series of potentially key factors/variables for detecting or probabilistically classifying the participants of the pilots. This also helps to create a series of risk patterns (of offender and victim) for the cybercrimes under consideration. However, the approach used in this task is based on a different mindset than the one used in Task 6.2 (based on Machine Learning predictions). In this task, an approach based on causality and Bayesian statistics is used.

More specifically, we have analysed the data collected in the RAYUELA pilots using Bayesian Networks. We have also been assisted by RAYUELA cyberbullying experts for proposing such network architectures. Subsequently, we perform a series of causal statistical analyses that help us identify key factors/drivers to determine the characteristics of potential victims and perpetrators. Finally, we note a series of comments and limitations on the techniques used and the data available so far, which make us cautious about the conclusions that can be drawn from the results.

It should be noted that the reliability of the results we can obtain depends on the cybercrime in question. In the case of cyberbullying, it is the only cybercrime considered for which we have a validated psychological questionnaire that players must answer [7]. This would be the data collected that is closest to a "ground truth" to serve as a validation/evaluation. In this way, the methods and conclusions drawn from the analysis of cyberbullying will be useful for the study of the other cybercrimes considered in RAYUELA.

Based on these findings, it appears that the variables collected through the RAYUELA serious game are promising for risk estimation. Although when looking for the strength of influence of multiple variables at the same time (i.e., multifactor analysis), the difference between the variables coming from the video game and those from profiling (demographic and psychological) narrows. It is important to exercise caution in interpreting the results due to the limited amount of data available for analysis, as well as the potential noise inherent in social science and video game data. Nevertheless, these initial results suggest that the RAYUELA serious game has the potential to be a valuable tool for social research purposes, highlighting the need for further exploration of its capabilities.

1. Introduction

This document summarises the data analysis performed in WP6 to analyse risk profiles and indicators of the considered cybercrimes in RAYUELA (cyberbullying, online grooming, cyberthreats, and fake news).

The proposed methodology consists in using Causal Graphical Models, also known as Bayesian Networks, to quantify the statistical and causal dependencies between variables that affect the occurrence of each cybercrime. Thus, we would identify the most relevant profiles and their indicators affecting each cybercrime among those considered. The obtained results will serve as a basis for proposing more effective interventions and regulatory actions based on scientific evidence.

Specifically, the goal is to use data extracted from the pilots performed during the RAYUELA project to train the models, find potential methodology limitations and define suitable metrics to validate the obtained results. Also, to make explicit the decisions and assumptions taken during this phase of the project. This report highlights the advantages of using Bayesian Networks, including combining insights from scientific literature, expert knowledge (WP1), and data gathered through the game (WP3 & WP5) to identify essential factors/drivers that determine whether an individual is a victim or an offender.

The RAYUELA serious game, due to ethical and legal constraints, has been designed to differentiate between participants with different bystander attitudes. However, through statistical prediction methods, such as those presented in deliverable D6.2 (Machine Learning methods) and D6.3 (Bayesian causal inference methods), we also aim to extract valuable information about the roles of the victims and perpetrators (whenever the perpetrators are minors).

Ultimately, the research questions we intend to answer in this report for each cybercrime considered are:

- **Research Question 1:** Which variables are most strongly related to the risk of suffering/committing each cybercrime?
- **Research Question 2:** What combinations of variables make it possible to construct meaningful risk profiles for each cybercrime?

The structure of the document is as follows: in Sec. 2, we explain the methodology applied to modelling with Bayesian Networks and the metrics used to validate the obtained results. In Sec. 3, we present the main results from the analysis of Cyberbullying. In Sec. 4, we present the main results from the analysis of Online Grooming. In Sec. 5, we present the main results from the analysis of Cyberthreats. In Sec. 6, we present the main results from the analysis of Fake News. We conclude in Sec. 7 with some preliminary results from the analysis but also point to limitations and future work.

2. Methodology

During the course of this work, we have taken an approach that diverges substantially from the techniques used in Task 6.2. While Task 6.2 relied on methods based on the Machine Learning mindset (i.e., a predictive-based mindset), the methodology utilised in the current task was founded on causality and Bayesian statistics principles. This shift in focus has proven more appropriate and effective when our objective was to explore and understand which factors have most significantly impacted a particular event [15-17].

The predictive mindset, which characterises traditional Machine Learning models, is often concerned with maximising the predictive power and pattern detection without necessarily understanding the underlying causal relationships. Consequently, the findings may be distorted by spurious correlations or biases in the data, leading to ambiguity and unexplainable relationships within complex systems. On the contrary, the causal approach seeks to provide a more profound understanding of how variables interact and influence one another. By leveraging Bayesian statistics, it becomes feasible to construct models that not only fit the observed data but also capture the underlying causal mechanisms that give rise to the data.

Specifically, our approach involved using a Probabilistic Graphical Model known as Bayesian Networks (BNs) [1]. These networks serve as a powerful tool to model the complexity of the relationships among various variables. The structure of such networks is represented as Directed Acyclic Graphs (DAGs), which encode the statistical and causal relationships between variables. The connections between the nodes in the DAG illustrate the conditional dependencies between the variables, and thus, encapsulate the joint probability distributions across them. BN fit perfectly for taking an occurred event and inferring the likelihood that each possible known cause was the contributing factor. Moreover, with this technique, we can ask the model counterfactual questions ("What would have happened if...?") and obtain a quantifiable and coherent answer with the available evidence, or even simulate interventions [2].

In a BN, nodes can represent two elements:

- **Observable Quantities:** Variables that can be directly observed or measured in the system.
- **Latent Quantities:** Hidden or unobserved variables that cannot be directly measured but may influence observed variables.

The use of this modelling technique is advantageous for several reasons. Firstly, BN allows us to combine insights from literature, expert knowledge, and data gathered through the game. It also motivates researchers to raise more questions about the data and to present assumptions and hypotheses, encouraging a fruitful debate explicitly. This approach is essential when dealing with sensitive social research topics [3]. Furthermore, it fits better with the probabilistic forecasting concept stipulated in the RAYUELA Grant Agreement.

To conduct the experiments, we used specialised BN software, GeNIe[11]. Using this tool, we can define a BN structure through a DAG, establish prior probabilities, and automatically fit the network parameters to the observed data. Once the BN has been constructed and trained, we *interrogate the model* (e.g., sensibility analysis, simulate intervention, etc.) to reveal which variables are the most determinant in each cybercrime.

Through the RAYUELA pilots conducted in schools with minors, we collected 1132 play sessions. Participants were between 12 and 16 years old (Mean=14.05, SD=1.38), where 57% identified themselves as males, 44%

[1] Bayes Fusion, GeNIe Modeler. https://www.bayesfusion.com/genie/

as females and 1% as non-binary. In Annex I we present an exploratory data analysis of all the data collected. The variables collected from the participants during these pilots are the following:

- Demographics: Age, gender, sexual orientation and migratory background.
- Technological: Daily hours spent on the Internet (for leisure).
- Psychological & Sociological Questionnaires: Social support (friends and significant other), family support, self-esteem, Big Five personality traits (agreeableness, neuroticism, extraversion, conscientiousness, openness to experience) and previous victimisation/offending.
    - A short version of the Big Five Inventory of personality traits questionnaire [4]
    - Rosenberg self-esteem scale [5]
    - The Multidimensional Scale of Perceived Social Support [6]
    - European Cyberbullying Intervention Project Questionnaire [7]
- Gameplay (inside RAYUELA's serious game): Answer to game decisions, response times, and final question about differences between their behaviour in the game and in reality («Have you played as you would behave in the real world? »).

The final question about the differences between their behaviour in the game and in reality, was proposed because we measured some variables both through the questionnaires and the situations of RAYUELA's serious game. Therefore, a self-reported estimate of the alignment of behaviours between reality and the game can be obtained for model calibration. Another indicator of the «honesty» of the players in their game responses is the gameplay reaction times. That is, we can reasonably assume that answers with extremely short reaction times are random choices of the player.

It should be noted that the analysis of cyberbullying (CB) is qualitatively different from the rest since we have a validated psychological questionnaire for this cybercrime [7], which is the closest measure to a 'ground truth'. The conclusions and lessons we draw from the CB analysis will allow us to validate this methodology to transfer it to the other cybercrimes for which, for ethical, legal or practical reasons, we cannot obtain a similar evaluation measure. However, bearing in mind that the conclusions we can draw from other cybercrimes will always have much greater uncertainty.

In the following, we detail the phases we followed in the experimental methodology. Firstly, we will address the generation of the BN structures (i.e., DAGs); secondly, the analyses and experiments carried out on the trained BNs, which will guide us to draw qualitative conclusions of profile matching and risk indicators.

2.1 Generation of BN structures

The determination of an appropriate BN structure (i.e., DAG) is a fundamental phase in any causal analysis process. This structure describes the specific relationships that exist between the variables of interest, and consequently the conclusions that can be drawn from the studies are intimately linked to the veracity and validity of the structure chosen.

It is important to understand that as the number of variables within a system increases, the number of viable BN structures follows an exponential growth pattern. This increasing complexity makes it practically infeasible to exhaustively evaluate and test all conceivable configurations. Such a computational and analytical challenge necessitates the use of causal discovery algorithms designed to navigate this vast search space.

There are numerous causal discovery algorithms that attempt to address this problem [8]. However, these algorithms often produce results of lower quality than the structures conceptualised and proposed by experts in the field. These shortcomings can manifest themselves in a variety of ways, such as lower accuracy in representing the underlying causal relationships or a lack of robustness to variations in the data.

In the previous deliverable D6.3, we performed a meticulous quantitative comparison between different BN structures, analysing them through various metrics and methodologies. The culmination of this comparative analysis yielded a decisive winner: the structure proposed by the domain experts.

For the current task, taking advantage of the lessons and knowledge gained from this exhaustive evaluation, we decided to directly use the structures provided by the RAYUELA experts. This decision was based both on empirical evidence supporting the superior performance of the structures proposed by the experts, and on practical considerations of efficiency and reliability. The utilisation of t-SNE (see Annex II) served as a complementary step to validate the suitability of the expert-defined structures provided by RAYUELA. This validation was underpinned by empirical evidence confirming the presence of latent variables in our dataset, aligning with the practical rationale of adopting structures that demonstrated both efficacy and dependability.

2.2 Variable Importance Analysis based on causality

After finalising the selection of BN structures, we conducted a series of causality-based experiments to interrogate the trained models, with the objective of drawing insightful conclusions about the underlying causes and mechanisms of these crimes.

The final questionnaire on cyberbullying victimisation and offending [7] serves as an evaluation measure ("ground truth"), providing us with empirical evidence and a firm basis for drawing conclusions that are closely aligned with the reality of this particular form of cybercrime.

In contrast, for the other types of cybercrime, we have used the same methodology as for CB, but with a significant limitation: the lack of an evaluation measure. This means that we lack a definitive standard against which to evaluate our conclusions. Thus, while we have endeavoured to apply rigorous analytical techniques, we must be cautious in interpreting our results, recognizing that it may not be possible to definitively determine the extent to which the conclusions drawn reflect the actual reality of the specific cybercrimes studied. The experimental analyses applied are detailed below.

2.2.1 Arrow Strength Analysis

This first analysis, also known as strength of influence, attempts to answer the question: *How strong is the causal influence of a cause on the variable of interest?* In other words, we will quantify the strength of influence of some variables on a variable of interest through all the possible paths through which this influence may propagate in the BN. The method to perform this analysis is based on the work of Koiter [10]. It consists mainly in measuring the similarity between several probability distributions of a variable of interest conditional on the states of the parent nodes. That is, we will marginalise the parent variables (i.e., simulate observed evidence) in the BN and then compare the different probability distributions obtained in the variable of interest.

There are numerous ways to compare probability distributions (i.e., quantify the difference between them). We have chosen the Jensen-Shannon distance [11], which is convenient because it is a symmetric distance measure, and it is also commonly used to perform this type of analysis [10]. In order to make the results easier to interpret, we will use their normalised version, whereby a value of 1 means total influence, and a value of 0 means no influence.

The Jensen-Shannon distance between two probability vectors *p* and *q* is defined as:

$$JSD(p\|q) = \sqrt{\frac{D(p\|m) + D(q\|m)}{2}}$$

Where *m* is the pointwise mean of *p* and *q*, and *D* is the Kullback-Leibler divergence [12]. The Kullback-Leibler divergence between a probability distribution *P* and a reference probability distribution *Q* is defined as:

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

Once the divergences of all the variables have been obtained and normalised, we can make a ranking of the variables that most influence the output variable. Those variables that obtain a greater Jensen-Shannon distance have the greatest power to influence the variable of interest.

2.2.2 Multi-Factor Marginalisation Analysis

This analysis also examines the importance of the variables in the model regarding a variable of interest. However, unlike in *2.2.1 Arrow Strength Analysis*, we are now interested in finding combinations of variables (i.e., multi-factor) that significantly affect the variable of interest. That is, that significantly changes the conditional probability distribution of the variable of interest after marginalising the values of the parent nodes. In addition, we will use this analysis to compare the relevance of variables coming from the gameplay with those from demographic variables or psychological questionnaires.

The variables that do not come from the gameplay are encompassed in the term "profiling." Player profiling consists of analysis or categorization using only static variables that are not (necessarily) directly related to gameplay [13]. Those are demographic variables (age, gender, sexual orientation, migratory background, daily hours of Internet use) and those collected through psychological questionnaires (social support, family support, self-esteem, previous CB victimisation, and previous CB offending).

The method used in this analysis consists of inserting multiple observations into the BN and recording the probabilities that have been updated. To do this, we will brute force all combinations of values for all the possible pieces of evidence for both cases (game questions and profiling). For example, 1 fixed piece of evidence could be *Age=14* or *Adventure 1 Question 3: Mathew Meme=Answer 2* (see Annex III to check the game decisions transcript). Through this analysis, we will also be able to determine risk profiles for each cybercrime, identifying the most relevant demographic and psychological variables observed in cybercrime perpetrators/victims.

3. Cyberbullying

This section describes the methodology used to analyse CB in adolescents through the RAYUELA video game sessions in the pilots. As described in the Methodology section, BNs are used to model and estimate the causal relationships and probabilistic dependencies between the variables under consideration.

3.1 Data Processing

Data Source and Filters:

The data source comes from the RAYUELA's serious game pilots conducted in schools, where adolescents face different situations related to online interaction. After applying filters to ensure data quality, **1147 valid records** were selected. Records with unfinished game sessions or those that did not provide relevant information for the analysis were eliminated (e.g., if they did not answer the final cyber-victimisation questionnaire or they did not play the corresponding adventures).

Gameplay Variables:

The video game variables reflect the decisions adolescents make in various online situations, such as sharing personal information, accepting friend requests, or sending photos to strangers, among others. These variables are fundamental to understanding adolescents' behaviour in virtual environments and their risk of committing/being victimised by BC. For this case, responses from **Adventures 1 and 3** are used, in which situations related to cyberbullying are developed.

Profiling (Socio-Demographic and Psychological) Variables:

In addition to the variables from the video game itself, socio-demographic data (such as age, gender or migratory background) and psychological test results of the participants, are considered. These factors can influence the probability of being a victim or offender of CB, making their inclusion essential for a more comprehensive and precise understanding of the phenomenon [18-21].

3.2 DAG Design

Figure 1 shows the BN architecture proposed by the CB experts for this case study, which is to analyse the variable of interest '**Previous CB Offending**'. As the number of nodes in the network increases, it becomes more difficult to visualise and interpret. Therefore, to facilitate this task, we have listed below the indicators/variables considered and to which ones they causally affect:

- Age affects:
    - Empathy
    - Previous CB Victimisation
    - Previous CB Offending
    - Daily Hours of Internet
- Gender affects:
    - Previous CB Victimisation
    - Previous CB Offending
- Sexual Orientation affects:
    - Previous CB Victimisation

- Migratory Background affects:
  - Previous CB Victimisation
- Daily Hours of Internet affects:
  - Previous CB Victimisation
  - Previous CB Offending
- Social Support affects:
  - Daily Hours of Internet
  - Empathy
  - Previous CB Victimisation
  - Previous CB Offending
- Family Support affects:
  - Daily Hours of Internet
  - Empathy
  - Previous CB Victimisation
  - Previous CB Offending
- Self-Esteem affects:
  - Previous CB Victimisation
  - Previous CB Offending
- Empathy affects:
  - Previous CB Offending
- Previous CB Victimisation affects:
  - Previous CB Offending
- Previous CB Offending affects:
  - “Honesty” Question
  - Game Answers
- “Honesty” Question affects:
  - Game Answers

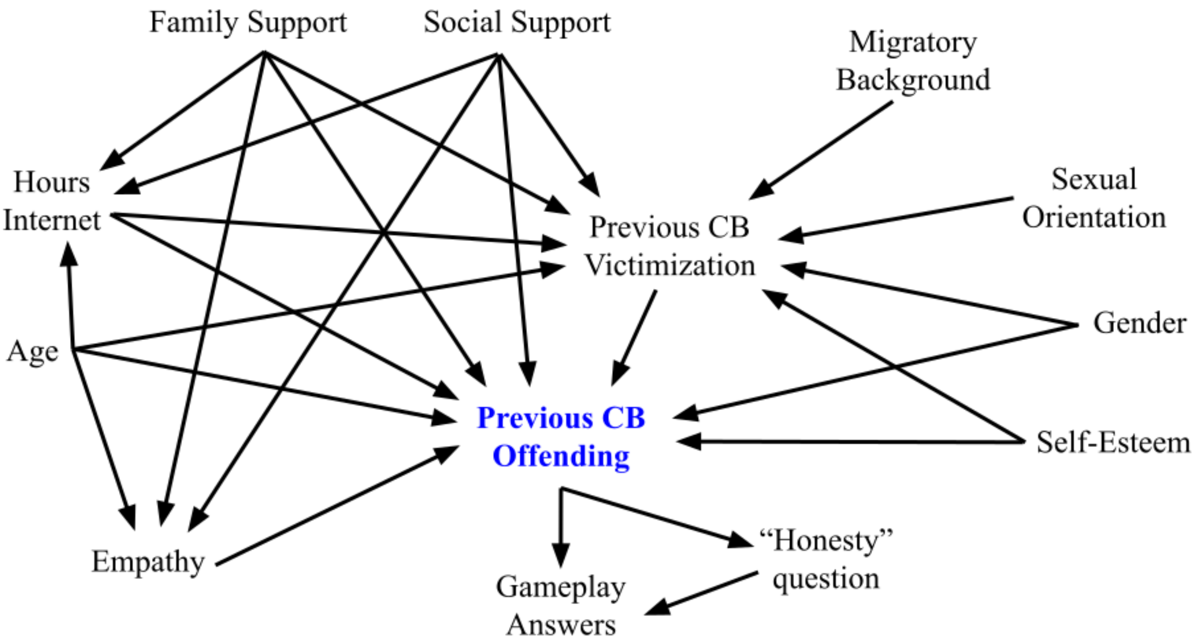


Figure 1. Structure of the Bayesian Network proposed by the CB experts of the RAYUELA project to analyse the variable of interest 'Previous CB Offending'. The variable of interest in this case study is highlighted in blue.

Figure 2 shows the BN architecture proposed by the CB experts for this case study, which is to analyse the variable of interest '**Previous CB Victimisation**. As the number of nodes in the network increases, it becomes more difficult to visualise and interpret. Therefore, to facilitate this task, we have listed below the indicators/variables considered and to which ones they causally affect:

- Age affects:
  - Previous CB Victimisation
  - Daily Hours of Internet
- Gender affects:
  - Previous CB Victimisation
- Sexual Orientation affects:
  - Previous CB Victimisation
- Migratory Background affects:
  - Previous CB Victimisation
- Daily Hours of Internet affects:
  - Previous CB Victimisation
- Social Support affects:
  - Daily Hours of Internet
  - Previous CB Victimisation
- Family Support affects:
  - Daily Hours of Internet
  - Previous CB Victimisation

- Self-Esteem affects:
  - Previous CB Victimisation
- Previous CB Victimisation affects:
  - "Honesty" Question
  - Game Answers
- "Honesty" Question affects:
  - Game Answers

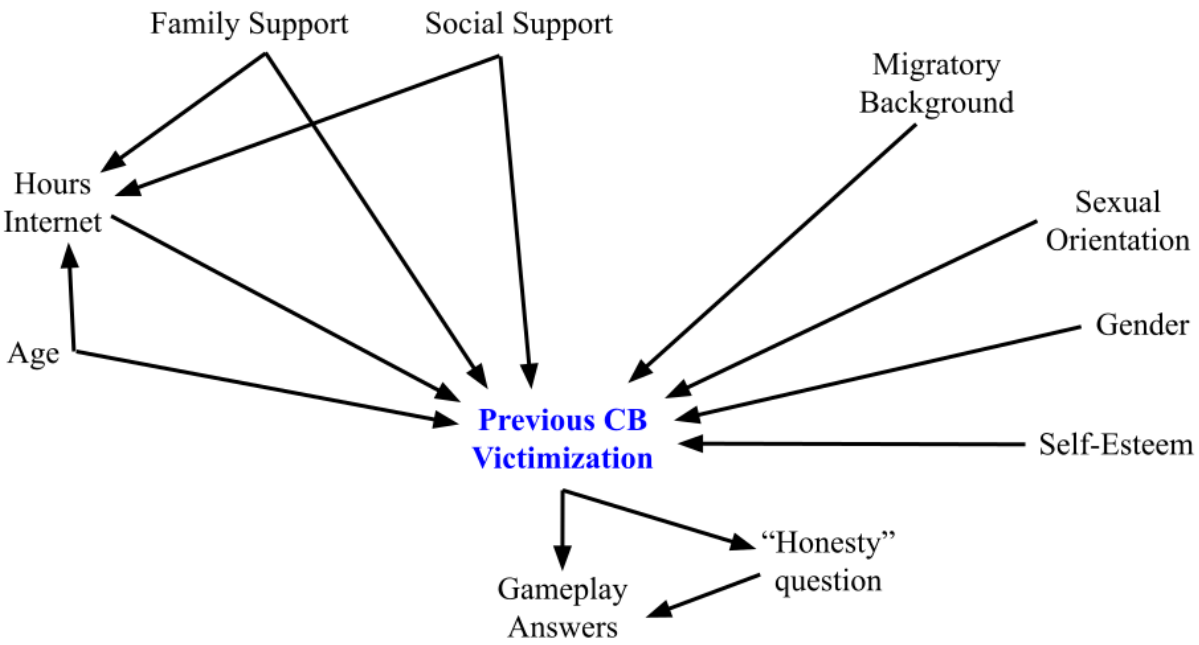


Figure 2. Structure of the Bayesian Network proposed by the CB experts of the RAYUELA project to analyse the variable of interest 'Previous CB Victimisation'. The variable of interest in this case study is highlighted in blue.

3.3 Experiment Results

Once we have created what we believe to be the best possible causal structure for this case study and the variables under consideration, we proceed with the quantitative experiments based on causality.

3.3.1 Arrow Strength Analysis

As we detailed in the Methodology section, the *strength of influence* analysis consists of simulating evidence in the BN for the variable of interest and comparing the probability distributions of each of the indicators/variables considered.

Firstly, we will focus on the case study of **CB offending** (Figure 1). Table 1 shows the results of this analysis on the variable of interest 'Previous CB Offending'.

Table 1. Strength of influence analysis of 'Previous CB Offending', on the selected Bayesian Network structure (proposed by CB experts). Normalised values: 1 means full influence and 0 means no influence. There is no standard value at which the influence is statistically significant, but in this case, we have indicated in green with an asterisk (*) those variables with values greater than 0.1. (See Annex III for more details on the content of the questions)

Indicators/Variable	Normalised Jensen-Shannon Distance
Adventure 3 Question 3: time overrun	0.20*
Adventure 3 Question 4: Pol Bullied	0.20*
Adventure 3 Question 5: Remind Matthew	0.18*
Adventure 1 Question 3: Matthew Meme	0.17*
Previous Victimisation	0.17*
Adventure 3 Question 1: Pirated Content	0.16*
Adventure 3 Question 7: Help Pol	0.12*
Adventure 3 Question 2: Pol Pola	0.11*
Adventure 3 Question 6: Talk Pol	0.08
"Honesty" question	0.07
Age	0.05
Gender	0.04
Daily Hours Internet	0.03
Family Support	0.02
Social Support	0.02
Self-Esteem	0.02
Adventure 1 Question 2: Sociable	0.01
Adventure 1 Question 1: Photo Sharing	0.01
Empathy	0.00

Secondly, we will focus on the case study of **CB Victimisation** (Figure 2). Table 2 shows the results of this analysis on the variable of interest 'Previous CB Victimisation'.

Table 2. Strength of influence analysis of Previous CB Victimisation', on the selected Bayesian Network structure (proposed by CB experts). Normalised values: 1 means full influence and 0 means no influence. There is no standard value at which the influence is statistically significant, but in this case, we have indicated in green with an asterisk (*) those variables with values greater than 0.1. (See Annex III for more details on the content of the questions)

Indicators/Variable	Normalised Jensen-Shannon Distance
Adventure 3 Question 5: Remind Matthew	0.13*
Adventure 3 Question 3: time overrun	0.12*
Adventure 3 Question 4: Pol Bullied	0.11*
Adventure 1 Question 3: Matthew Meme	0.09
Adventure 3 Question 1: Pirated Content	0.08
Adventure 3 Question 7: Help Pol	0.08
Daily Hours Internet	0.07
Adventure 3 Question 2: Pol Pola	0.07
"Honesty" question	0.06
Sexual Orientation	0.06
Migratory Background	0.06
Age	0.05
Social Support	0.05

Gender	0.05
Family Support	0.05
Self-Esteem	0.03
Adventure 1 Question 1: Photo Sharing	0.02
Adventure 3 Question 6: Talk Pol	0.01
Adventure 1 Question 2: Sociable	0.00

3.3.2 Multi-Factor Marginalisation Analysis

As we detailed in the Methodology section, this analysis also examines the importance of the variables in the model. However, unlike *Arrow Strength Analysis*, we are now interested in finding combinations of variables (i.e., multi-factor) that significantly change the conditional probability of the variable of interest. In addition, we will use this analysis to compare the relevance of variables coming from the gameplay with those from demographic variables or psychological questionnaires.

First, we have analysed the CB Offending case study. Figure 3 shows the maximum conditional probability of a positive response to the variable of interest '**Previous CB Offending**' as a function of the amount of evidence inserted in the BN. Results are presented for both data sources (game questions and profiling).

Also shown are two lines marking significant values of the conditional probability of the variable of interest compared to the prior probability distribution, which was set to 0.1 before training the BN parameters, with an effective sample size of 2 (soft prior). Using the Jeffreys scale [14] for comparing odds ratios, a ratio between $10^{1/2}$ and 10 is interpreted as a substantial difference. A ratio between 10 and $10^{3/2}$ is a strong difference. In our case, with 0.1 prior probability, this would occur with posterior probabilities of ~0.26 and ~0.53, respectively (Equation 1). Although, it is essential to remember that as the number of fixed pieces of evidence increases, the number of players who meet these criteria (i.e., probability of evidence) will decrease.

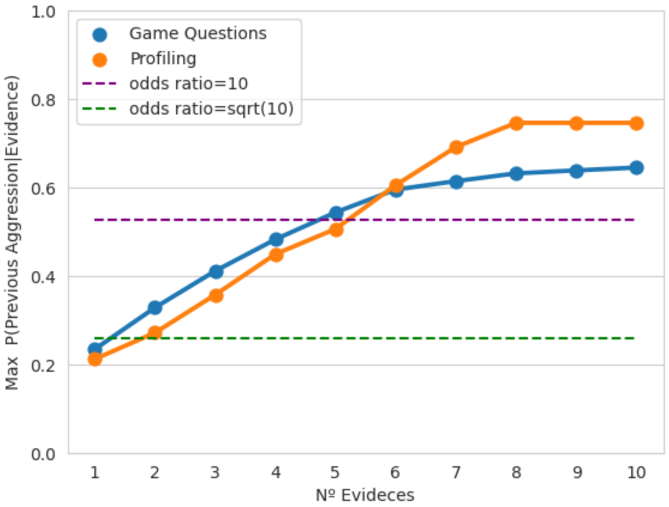


Figure 3. By performing a multi-factor analysis, we can find the combinations of variables that cause a greater increase in the conditional probability of the outcome (*Previous CB Offending*). The following figure is obtained by finding the maximum conditional probability obtained by setting different numbers of combinations of evidence (from 1 to 10). This is done, on the one hand, for the variables obtained through the game questions and, on the other hand, for the profiling variables. The figure also shows the conditional probabilities corresponding to the relevant thresholds according to Jeffreys' criterion [14] calculated in Equation 1.

$$BF = \frac{\text{prior odds ratio}}{\text{posterior odds ratio}} = \frac{0.9/0.1}{(1-X)/X}$$

$$BF = 10^{1/2} \text{ (Substantial evidence)} \Rightarrow X \approx 0.26$$

$$BF = 10 \text{ (Strong evidence)} \Rightarrow X \approx 0.53$$

Equation 1. Calculating the probability thresholds according to the Jeffreys criterion [14] and the selected prior

From 6 fixed pieces of evidence, the profile variables exceed the second threshold ~0.53. Further analysing this case, Figure 4 shows the number of observations of the risk profiles' most common shared profiling characteristics. We define a *risk profile* as one with a posterior probability (Previous CB Offending = True) greater than or equal to 0.26. These risk profiles' top shared profiling characteristics are as follows: *Previous CB victimisation=True, gender=Male, Social Support = High, and Family Support = High*.

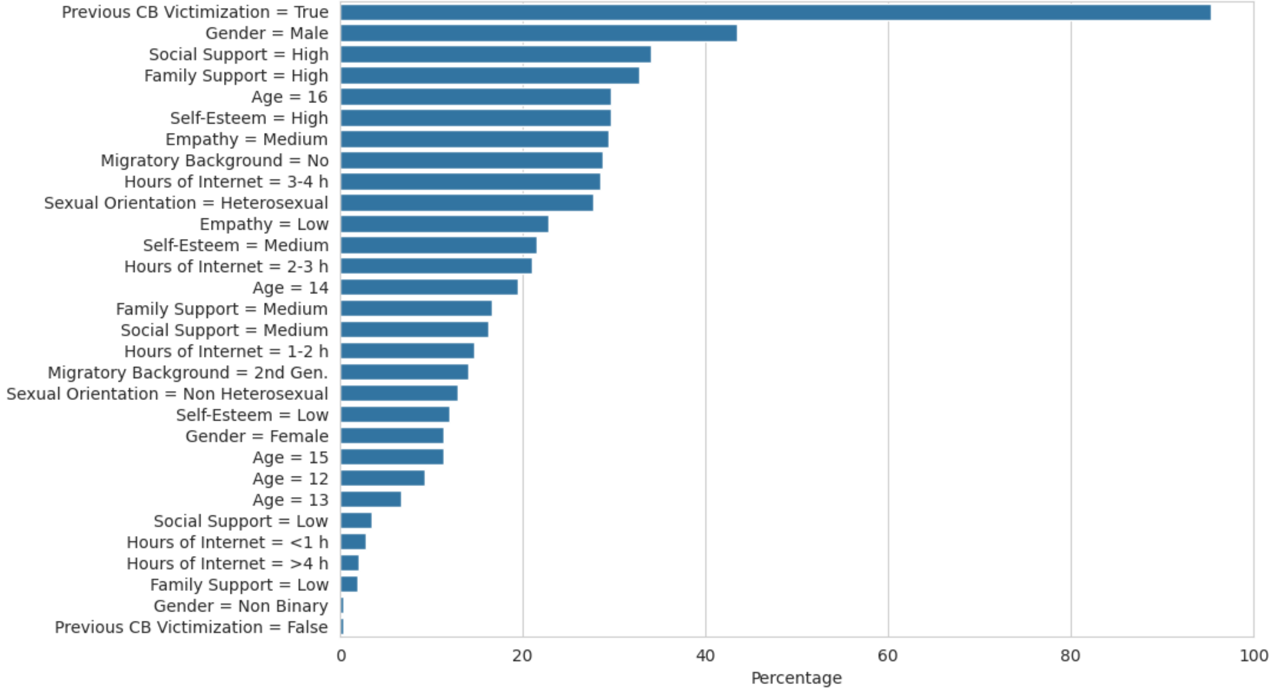


Figure 4. By performing a multi-factor analysis, we can find the combinations of variables (i.e., profiles) that the model has learned are "risky" for committing cyberbullying. The figure shows a count of the number of times a particular value of a variable appears in the identified risk profiles. In this case, we analyse the profiles obtained by setting exactly 6 variables, since from this number onwards, we begin to obtain risky profiles with a relevant odds ratio according to Jeffreys' criterion [14]. The value *Previous CB Victimisation = True* appears in more than 90% of the risk profiles when setting 6 evidences.

Second, we have analysed the CB Victimisation case study. Figure 5 shows the maximum conditional probability of a positive response to the variable of interest **'Previous CB Victimisation'** as a function of the amount of evidence inserted in the BN. Results are presented for both data sources (game questions and profiling). The same priors have been used as for the previous case, so the probability thresholds obtained from Equation 1 are identical. In this case it seems that the combinations of game variables are still more relevant than those of profiling (Figure 5). Although the difference is slightly greater than in the previous case.

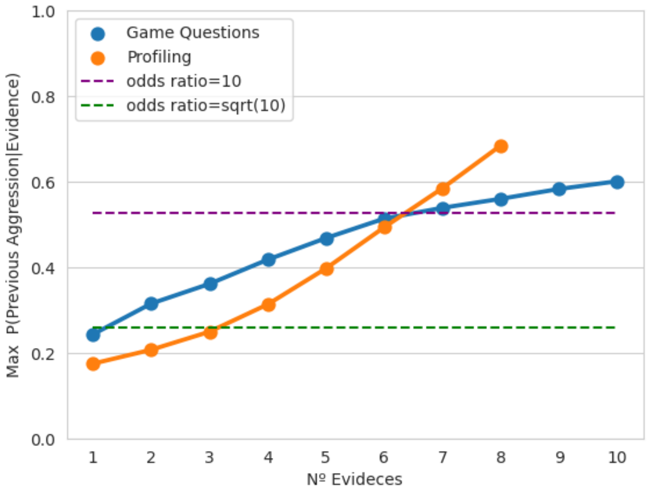


Figure 5. By performing a multi-factor analysis, we can find the combinations of variables that cause a greater increase in the conditional probability of the outcome (*Previous CB Victimisation*). The following figure is obtained by finding the maximum conditional probability obtained by setting different numbers of combinations of evidence (from 1 to 10). This is done, on the one hand, for the variables obtained through the game questions and, on the other hand, for the profiling variables. The figure also shows the conditional probabilities corresponding to the relevant thresholds according to Jeffreys' criterion [14] calculated in Equation 1.

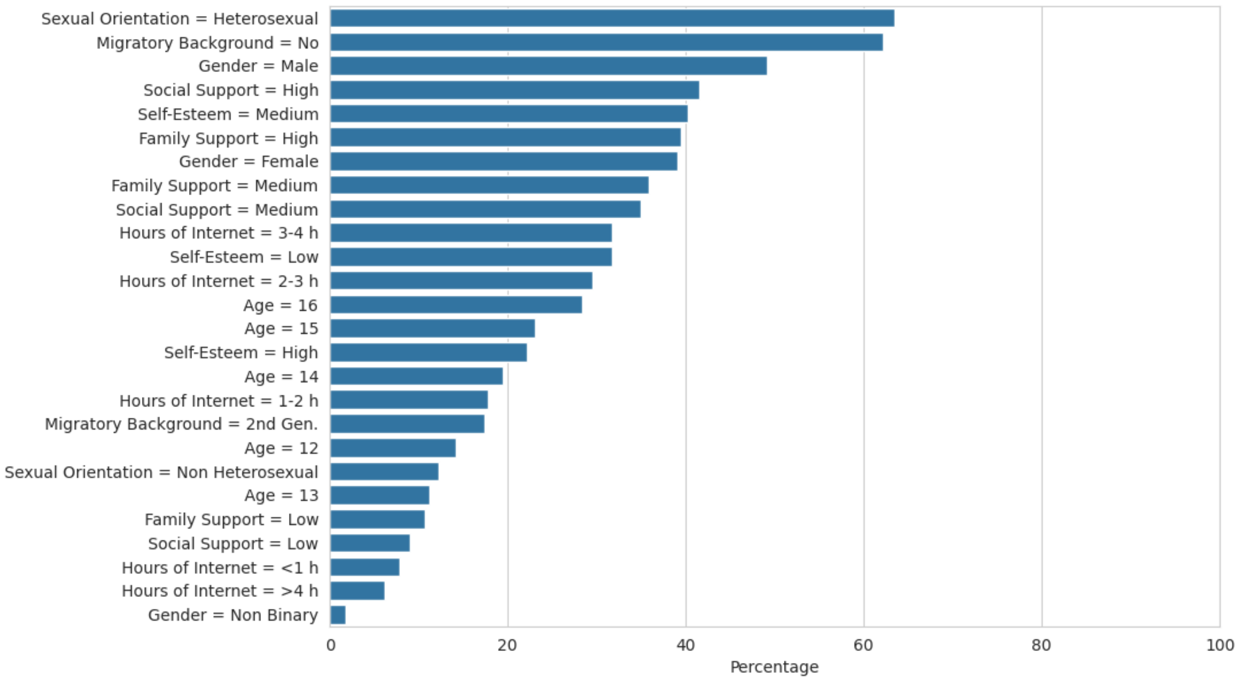


Figure 6. By performing a multi-factor analysis, we can find the combinations of variables (i.e., profiles) that the model has learned are "risky" for suffering cyberbullying. The figure shows a count of the number of times a particular value of a variable appears in the identified risk profiles. In this case, we analyse the profiles obtained by setting exactly 7 variables, since from this number onwards, we begin to obtain risky profiles with a relevant odds ratio according to Jeffreys' criterion [14].

From 7 fixed pieces of evidence, the profile variables exceed the second threshold ~0.53. Further analysing this case, Figure 6 shows the number of observations of the risk profiles' most common shared profiling characteristics. We define a *risk profile* as one with a posterior probability (Previous CB Offending = True) greater than or equal to 0.26. These risk profiles' top shared profiling characteristics are as follows: *Sexual Orientation = Heterosexual, Migratory Background = No, Gender = Male, and Social Support = High.*

3.4 Discussion

Based on the assumption that the proposed BN structures posited by experts in the field are accurate, the outcomes from the experiments provide compelling insights into the relationship between certain variables and patterns of behaviour related to CB offending and victimisation.

Initially, the strength of each variable has been studied separately to explain the outcome of the variables of interest, i.e., previous incidents of CB delinquency and victimisation. The study data clearly suggest that these video game-related questions have greater explanatory power compared to the profiling variables. However, an interesting exception was observed in the case of the variable "Previous CB Victimisation" when related to CB offending. Individuals who have previously been victims of CB exhibited a markedly higher propensity to commit CB offences. This relationship merits further study although it was one of the predictions of the work developed by WP1 to understand in depth the cybercrimes considered.

Second, multiple combinations of variables and their strength of influence were analysed. In this case the clear distinction between video game variables and profile variables became less pronounced. This multifactorial analysis allowed us to identify unique profiles that could be predictive of risk factors associated with CB delinquency and victimisation.

For CB offending, the multi-factor profiling revealed a convergence of characteristics that the risk profiles commonly share. Individuals who were previously victimised by CB (Previous CB victimisation=True), predominantly males, with robust levels of social and family support, emerged as more prone to becoming offenders. In the case of CB victimisation, the study identified certain traits such as heterosexual orientation, absence of a migratory background, male gender, and high levels of social support as common among the victims.

4. Online Grooming

This document describes the methodology used to analyse the probability of "Online Grooming" occurrence in adolescents through the RAYUELA video game sessions in the pilots. As described in the Methodology section, BNs are used to model and estimate the causal relationships and probabilistic dependencies between the variables under consideration.

4.1 Data Processing

Data Source and Filters:

The data source comes from the "Online Adventures" video game, where adolescents face different situations related to online interaction. After applying filters to ensure data quality, 1301 valid records were selected. It is important to eliminate records with ambiguous responses or those that do not provide relevant information for the analysis.

Gameplay Variables:

The selected variables primarily come from **Adventure 2 and 5**.

- [Adventure 2] Registration name: Participants provide information about their name, year of birth, favourite music band, or another beloved famous/TV/book character.
- [Adventure 2] Registration Profile time: Participants must decide to have a public/private profile in the social network.
- [Adventure 2] Registration place: Participants disclose details about their city, neighbourhood, school, country…
- [Adventure 2] Registration profile photo: Participants choose their profile photo among a photo of them, a photo with friends or something random from the Internet.
- [Adventure 2] Use PC: Participants choose between view the received messages or check the profile of the sender.
- [Adventure 2] Friend Request: Participants decide whether to accept or reject a friend request and may choose to check the photographer's profile.
- [Adventure 2] Send photos: Participants decide whether to send or not send photos to the contact.
- [Adventure 2] More photos: Participants decide whether to send naked photos or not send them.
- [Adventure 2] More & more: Participants decide whether to send more photos.
- [Adventure 2] Ask Help: Participants decide to ask for help to Mary.
- [Adventure 2] Close case: Participants decide to check the profile or not.
- [Adventure 2] Tell Parents: Participants decide to tell what happened.
- [Adventure 2] Block profile: Participants decide whether to block the profile.
- [Adventure 5] Secret relationship: Participants express their opinion about Sheila's relationship
- [Adventure 5] Biology paper: Participants choose preference to meet online or in person
- [Adventure 5] Talk to Sheila: Participants choose whether to talk to Sheila.

Profiling (Socio-Demographic and Psychological) Variables:

In addition to the variables in the video game, socio-demographic data, such as age and gender of the participants, are considered. These factors can influence the probability of falling victim to "Online Grooming," making their inclusion essential for a more comprehensive understanding of the phenomenon.

4.2 DAG Design

The construction of a Bayesian network is a crucial part of this methodology as it allows modelling the causal relationships between variables. The network structure was designed with input from experts in the field and previous knowledge of "Online Grooming," ensuring that relevant variables are included and significant interactions are captured.

The EM algorithm is used to estimate the parameters of the network given the established structure and priors. The contribution of experts is key in defining the priors, which reflect the prevalence of "Online Grooming" in adolescents. Additionally, the expertise of the experts validates and refines the model, ensuring that the inferences obtained are reliable and adequately represent the studied phenomenon.

Figure 7 shows the BN architecture proposed by the experts for this case study, which is to analyse the variable of interest '**Online Grooming Victimisation Risk**'. As the number of nodes in the network increases, it becomes more difficult to visualise and interpret. Therefore, to facilitate this task, we have listed below the indicators/variables considered and to which ones they causally affect:

- Age affects:
    - Daily Hours of Internet
    - Online Grooming Victimisation Risk
- Gender affects:
    - Online Grooming Victimisation Risk
- Sexual Orientation affects:
    - Online Grooming Victimisation Risk
- Migratory Background affects:
    - Online Grooming Victimisation Risk
- Self-Esteem affects:
    - Online Grooming Victimisation Risk
- Daily Hours of Internet affects:
    - Online Grooming Victimisation Risk
- Social Support affects:
    - Daily Hours of Internet
    - Online Grooming Victimisation Risk
- Family Support affects:
    - Daily Hours of Internet
    - Online Grooming Victimisation Risk
- Introversion (BF) affects:
    - Daily Hours of Internet
    - Online Grooming Victimisation Risk
- Agreeableness (BF) affects:
    - Online Grooming Victimisation Risk
- Openness to Experience (BF) affects:
    - Online Grooming Victimisation Risk
- Neuroticism (BF) affects:

    - Online Grooming Victimisation Risk
- "Honesty" Question affects:
    - Game Answers
- Online grooming Victimisation Risk affects:
    - Game Answers

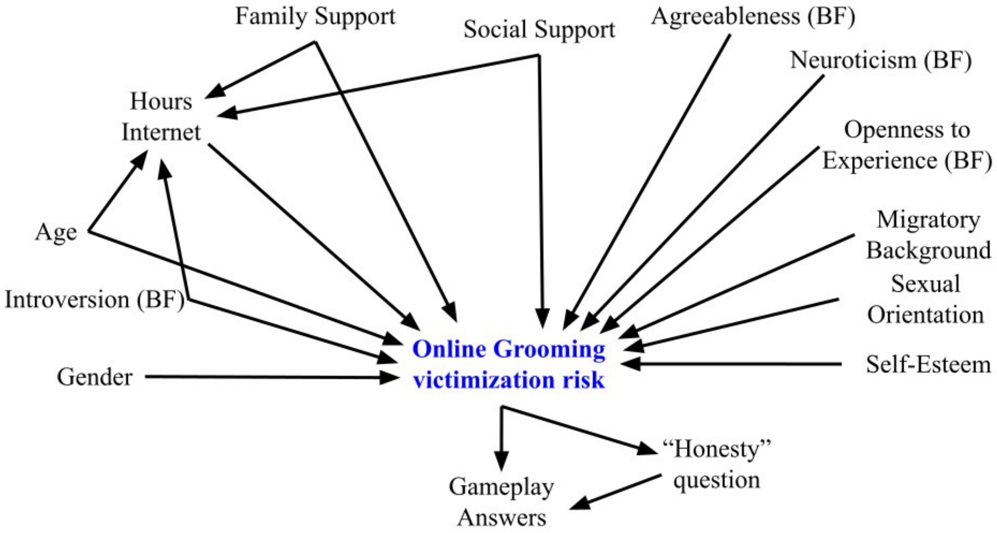


Figure 7. Structure of the Bayesian Network proposed by the experts of the RAYUELA project to analyse the variable of interest 'Online Grooming victimisation risk'. The variable of interest in this case study is highlighted in blue.

4.3 Experiment Results

Once we have created what we believe to be the best possible causal structure for this case study and the variables under consideration, we proceed with the quantitative experiments based on causality.

4.3.1 Arrow Strength Analysis

As we outlined in the Methodology section, the strength of influence analysis involves simulating evidence in the BN for the variable of interest and comparing the probability distributions of each of the considered indicators/variables. Table 3 presents the results obtained from this analysis, using the Jensen-Shannon divergence between the probability distributions conditioned on the "Online Grooming risk" variable.

Table 3. Strength of influence analysis of OG victimisation risk, on the selected Bayesian Network structure (proposed by experts). Normalised values: 1 means full influence and 0 means no influence. There is no standard value at which the influence is statistically significant, but in this case, we have indicated in green with an asterisk (*) those variables with values greater than 0.1. (See Annex III for more details on the content of the questions)

Indicators/Variable	Normalised Jensen-Shannon Distance
Adventure 2 Question 4 Place	0.39*
Adventure 2 Question 8 Friend Request	0.37*
Adventure 2 Question 9 Photos:	0.35*
Adventure 2 Question 3 Professional Type	0.34*
Adventure 2 Question 5 Profile Photo	0.25*
Adventure 5 Question 1 Secret	0.20*

Adventure 2 Question 15 Block Profile	0.19*
Adventure 2 Question 1 Name	0.16*
Adventure 5 Question 2 Biology	0.16*
Adventure 2 Question 14 Tell Parents	0.16*
Adventure 2 Question 12 Ask Help	0.14*
Adventure 2 Question 11 more & more	0.11*
Adventure 2 Question 7 Use PC	0.11*
Adventure 2 Question 13 Close Case	0.08
Honesty	0.055
Adventure 5 Question 3 Sheila	0.01
Adventure 2 Question 10 Photos NK	0.01
Gender	0.01
Family Support	0.00
Migratory Background	0.00
Social Support	0.00
Openness to Experience	0.00
Neuroticism	0.00
Sexual Orientation	0.00
Hours of Internet Usage	0.00

4.3.2 Multi-Factor Marginalisation Analysis

As we elaborated in the Methodology section, this analysis also explores the significance of variables within the model. However, unlike the Arrow Strength Analysis, our focus here lies in identifying combinations of variables (i.e., multifactorial) that substantially alter the conditional probability of the target variable. Furthermore, we will utilize this analysis to compare the significance of variables derived from gameplay-related questions with those originating from demographic factors or psychological questionnaires.

Figure 8 displays the maximum conditional probability of a positive response to the "Online Grooming Risk" variable, with the extent of inserted evidence varying across the Bayesian Network. These results are presented for both data sources (game questions and profiling).

Two lines are also depicted, demarcating notable values of the conditional probability for the target variable concerning the prior probability distribution. The initial distribution was set to 0.95 for a positive response and 0.05 for a negative response before the BN parameters were trained. Using the Jeffreys scale [14] for comparing odds ratios, a substantial difference between prior and posterior probabilities is obtained at a posterior probability of ~0.14, and a strong difference at ~0.34 (Equation 2).

$$
\begin{gathered}
BF=\frac{\text{prior odds ratio}}{\text{posterior odds ratio}}=\frac{0.95/0.05}{(1-X)/X} \\
BF=10^{1/2}(\text{ Substantial evidence }) \Rightarrow X \approx 0.14 \\
BF=10(\text{ Strong evidence }) \Rightarrow X \approx 0.34
\end{gathered}
$$

Equation 2. Calculating the probability thresholds according to the Jeffreys criterion [14] and the selected prior

Nonetheless, it is important to bear in mind that as the number of fixed evidences points increases, the count of players conforming to these criteria (i.e., evidence probability) will decrease.

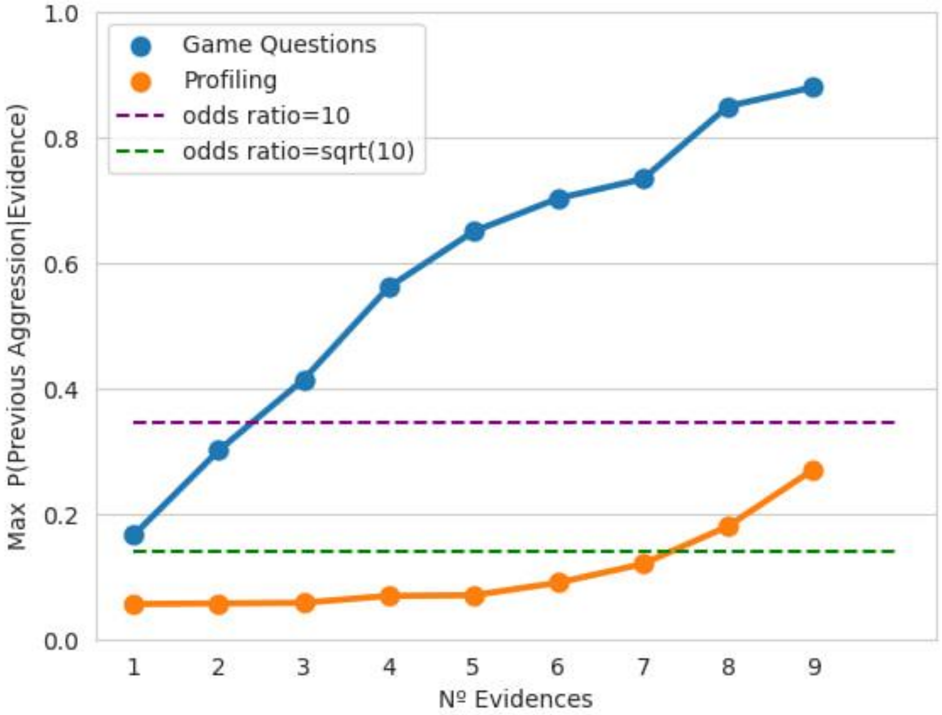


Figure 8. By performing a multi-factor analysis, we can find the combinations of variables that cause a greater increase in the conditional probability of the outcome (*OG victimisation risk*). The following figure is obtained by finding the maximum conditional probability obtained by setting different numbers of combinations of evidence. This is done, on the one hand, for the variables obtained through the game questions and, on the other hand, for the profiling variables. The figure also shows the conditional probabilities corresponding to the relevant thresholds according to Jeffreys' criterion [14] calculated in Equation 2.

Our findings reveal that, in order to achieve effective predictive power, we require a minimum of 8 demographic pieces of evidence, surpassing the initial threshold of ~0.14. However, if we look at the curve of the game questions, we only need 3 questions to pass the second threshold (~0.34). These results suggest that in the case of OG-related video game questions, profiling based solely on demographic/psychological variables is highly challenging.

Figure 9, presented within this context, offers a visual representation of the frequency of observations associated with the most prevalent shared attributes among profiles at risk. Within our study of OG, a 'risk profile' is defined as one with a posterior probability (Online Grooming Risk = True) equal to or greater than 0.34. Noteworthy characteristics of these risk profiles include lower levels of honesty, male gender, strong family support, moderate neuroticism, high social support, moderate conscientiousness, and low agreeableness.

It is crucial to highlight that our analysis encompasses a wide range of sociodemographic and psychological variables, including honesty, gender, age, and family support, among others. While these factors may exert a somewhat lesser influence compared to those tied to online behaviour, they still hold significant sway within the broader context of our study.

Additionally, our observations reveal that variables associated with in-game behaviour demonstrate remarkable sensitivity, requiring less data to surpass the upper threshold of probability. This phenomenon may be attributed to the network's structural dynamics and the direct impact of these variables on the latent target variable. This finding reinforces a recurring theme we have discerned throughout our analyses – that, regardless of predisposing demographic risk factors, the ultimate determinant of victimization lies in an individual's response to online situations.

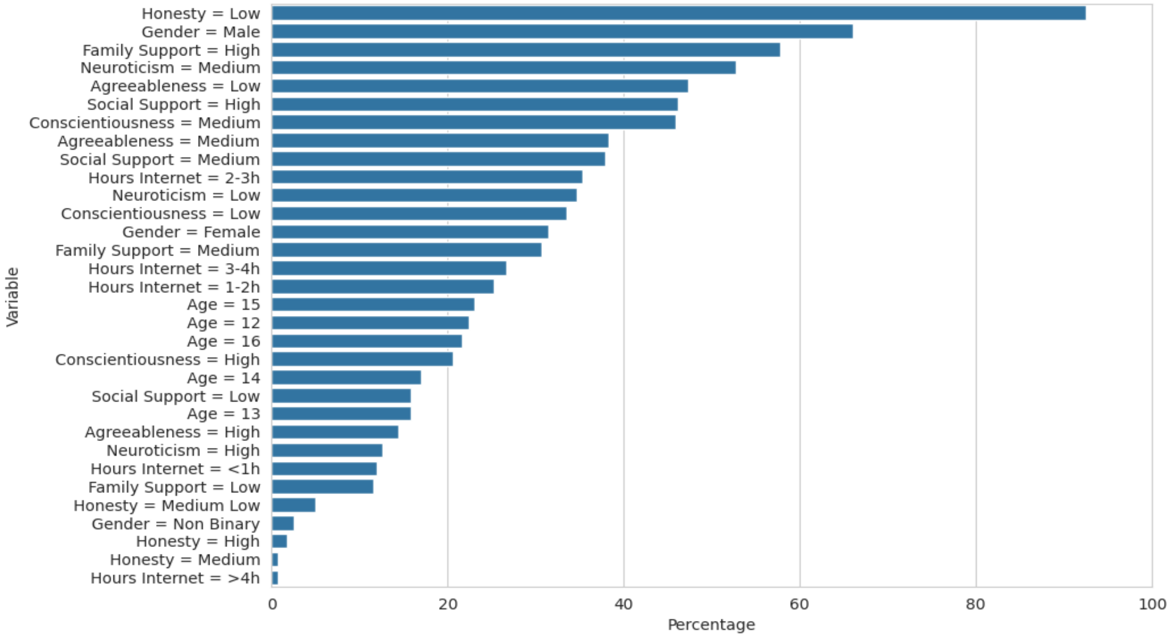


Figure 9. By performing a multi-factor analysis, we can find the combinations of variables (i.e., profiles) that the model has learned are "risky" for suffering OG. The figure shows a count of the number of times a particular value of a variable appears in the identified risk profiles. In this case, we analyse the profiles obtained by setting exactly 9 variables, since from this number onwards, we begin to obtain risky profiles with a relevant odds ratio according to Jeffreys' criterion [14].

4.4 Discussion

Based on the assumption that the proposed BN structures posited by experts in the field are accurate, the outcomes from the experiments provide compelling insights into the relationship between certain variables and patterns of behaviour related to OG victimisation risk.

Initially, the strength of each variable has been studied separately to explain the outcome of the variables of interest, i.e., OG victimisation risk. The study data clearly suggest that these video game-related questions have greater explanatory power compared to the profiling variables. Particularly, decisions related to registration place, acceptance of friend requests, photo sending, and profile type also play a significant role in the probability of falling victim to this cybercrime. By considering these variables in future analyses and preventive measures, more effective solutions can be designed to address this important issue.

Secondly, the multifactorial analysis allows us to study the combined effects of different levels of each discrete variable, representing the various possible combinations of responses to the questions posed by the video game, as well as the combination of profile variable values. Our findings suggest that, in this case, the game questions are more relevant and enable a more accurate discrimination compared to the profiling variables.

However, in this case, unlike the CB case, there is no "ground truth" of the OG victimisation risk variable. That is, now the variable of interest is latent. This implies that the methodology used can be interpreted as a Bayesian unsupervised clustering. That is, using the variables considered, we try to identify two distinct groups of people in the data. However, there is no guarantee that these groups actually correspond to minors at greater or lesser risk suffering the cybercrime under consideration. This also means that the values obtained in the experiments could be exaggerated or distorted.

5. Cyberthreats

This section describes the methodology used to analyse Cyberthreats in adolescents through the RAYUELA video game sessions in the pilots. As described in the Methodology section, BNs are used to model and estimate the causal relationships and probabilistic dependencies between the variables under consideration. Cyberthreats encompass various malicious activities conducted through digital channels that pose significant risks to young internet users. As adolescents are increasingly active online, they become more vulnerable to phishing scams, identity theft, online predators, and exposure to inappropriate content. Understanding these threats and their impact on teenagers is crucial for designing effective prevention strategies and fostering a safer digital environment.

5.1 Data Processing

Data Source and Filters:

The data source comes from the RAYUELA's serious game pilots conducted in schools, where adolescents face different situations related to online interaction. After applying filters to ensure data quality, **716 valid records** were selected. The game presents simulated scenarios of Cyberthreats, reflecting adolescents' decision-making processes in various online situations.

Gameplay Variables:

The selected variables primarily come from **Adventure 4**, with some contributions from Adventures 2 and 3.

- [Adventure 4] Phishing Email: In this scenario, adolescents must decide whether to act upon a suspicious email containing potential threats to their online accounts.
- [Adventure 4] New Password: Participants make choices regarding their account passwords, influencing their vulnerability to identity theft and unauthorised access.
- [Adventure 4] My Account Stolen: Participants make choices regarding their account theft.
- [Adventure 4] Other Account Stolen: Participants make choices regarding other people's account theft.
- [Adventure 3] Pirated Content: This variable examines adolescents' responses to downloading pirated content, exposing them to risks related to illegal online activities.
- [Adventure 2] Registration Password: Participants decide on the strength and complexity of their passwords during registration, affecting their susceptibility to hacking and unauthorised access.

Profiling (Socio-Demographic and Psychological) Variables:

In addition to the variables from the video game itself, socio-demographic data (such as age, gender or migratory background) and psychological test results of the participants, are considered. These factors can influence the probability of being a victim of a cyberattack, making their inclusion essential for a more comprehensive and precise understanding of the phenomenon.

5.2 DAG Design

The central variable of interest, "CT_risk," represents the **latent variable** indicating the risk of falling victim to Cyberthreats. All other variables in the network, including the decisions made by adolescents in response to various scenarios, are considered children or parent nodes of the "CT_risk" node. This design reflects how each decision is influenced by the overall risk level posed by Cyberthreats.

The combination of video game variables and socio-demographic information allows us to build a comprehensive model that captures the complexity of cyberthreats' interactions and their impact on adolescents. This DAG design ensures that the analysis considers both behavioural decisions and individual characteristics when assessing the risk of encountering cyberthreats.

Figure 10 shows the BN architecture proposed by the experts for this case study, which is to analyse the variable of interest '**Cyber Threats Risk**'. As the number of nodes in the network increases, it becomes more difficult to visualise and interpret. Therefore, to facilitate this task, we have listed below the indicators/variables considered and to which ones they causally affect:

- Age affects:
  - Daily Hours of Internet
  - Cyberthreats Risk
- Gender affects:
  - Cyberthreats Risk
- Daily Hours of Internet affects:
  - Cyberthreats Risk
- Social Support affects:
  - Daily Hours of Internet
  - Cyberthreats Risk
- Family Support affects:
  - Daily Hours of Internet
  - Cyberthreats Risk
- Agreeableness (BF) affects:
  - Cyberthreats Risk
- Neuroticism (BF) affects:
  - Cyberthreats Risk
- Conscientiousness (BF) affects:
  - Cyberthreats Risk
- "Honesty" Question affects:
  - Game Answers
- Cyberthreats Risk affects:
  - Game Answers

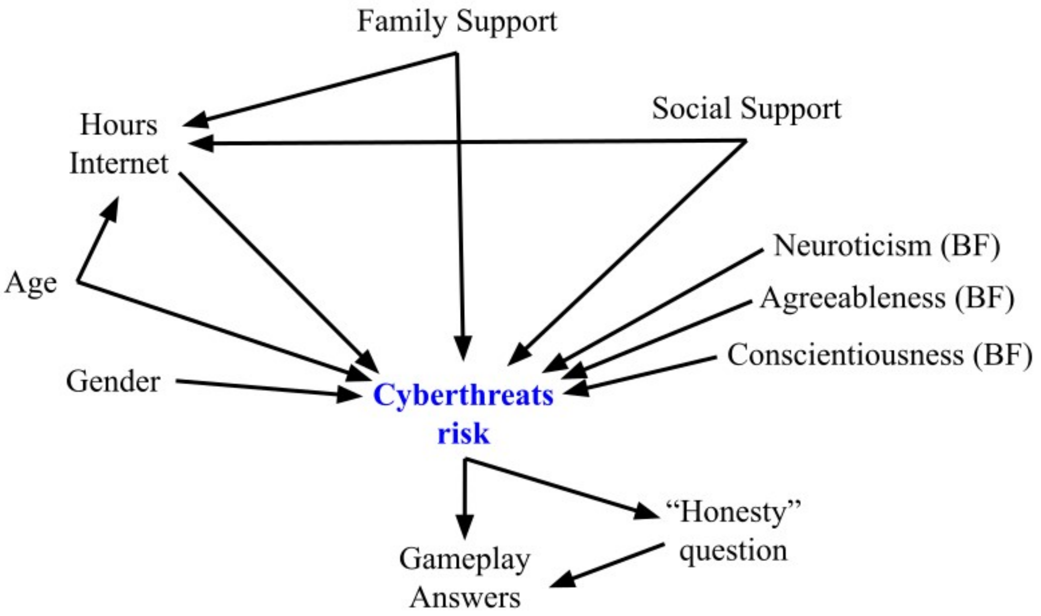


Figure 10. Structure of the Bayesian Network proposed by the experts of the RAYUELA project to analyse the variable of interest 'Cyberthreats risk'. The variable of interest in this case study is highlighted in blue.

5.3 Experiment Results

Once we have created what we believe to be the best possible causal structure for this case study and the variables under consideration, we proceed with the quantitative experiments based on causality.

5.3.1 Arrow Strength Analysis

As we detailed in the Methodology section, the *strength of influence* analysis consists of simulating evidence in the BN for the variable of interest and comparing the probability distributions of each of the indicators/variables considered. Table 4 shows the results obtained from the analysis, using the Jensen-Shannon divergence between the probability distributions conditioned on the "Cyberthreats risk" variable.

Table 4. Strength of influence analysis of Cyberthreats risk victimisation, on the selected Bayesian Network structure (proposed by experts). Normalised values: 1 means full influence and 0 means no influence. There is no standard value at which the influence is statistically significant, but in this case, we have indicated in green with an asterisk (*) those variables with values greater than 0.1. (See Annex III for more details on the content of the questions)

Indicators/Variable	Normalised Jensen-Shannon Distance
Adventure 4 Question 2: New Password	0.75*
Adventure 2 Question 2: Registration Password	0.45*
Adventure 4 Question 3: My Account Stolen	0.37*
Adventure 4 Question 1: Phishing	0.30*
Adventure 4 Question 4: Other Account Stolen	0.23*
Adventure 3 Question 1: Pirated Content	0.23*
"Honesty" question	0.20*
Gender	0.01
Daily Hours Internet	0.01
Age	0.01

Neuroticism (BF)	0.00
Social Support	0.00
Agreeableness (BF)	0.00
Conscientiousness (BF)	0.00
Family Support	0.00

5.3.2 Multi-Factor Marginalisation Analysis

As we detailed in the Methodology section, this analysis also examines the importance of the variables in the model. However, unlike *Arrow Strength Analysis*, we are now interested in finding combinations of variables (i.e., multi-factor) that significantly change the conditional probability of the variable of interest. In addition, we will use this analysis to compare the relevance of variables coming from the gameplay with those from demographic variables or psychological questionnaires.

Figure 11 shows the maximum conditional probability of a positive response to the variable of interest '**Cyberthreats risk**' as a function of the amount of evidence inserted in the BN. Results are presented for both data sources (game questions and profiling).

Also shown are two lines marking significant values of the conditional probability of the variable of interest compared to the prior probability distribution, which was set to 0.3 before training the BN parameters, with an effective sample size of 2 (soft prior). Using the Jeffreys scale [14] for comparing odds ratios, a ratio between $10^{1/2}$ and 10 is interpreted as a substantial difference. A ratio between 10 and $10^{3/2}$ is a strong difference. In our case, with 0.3 prior probability, this would occur with posterior probabilities of ~0.57 and ~0.81, respectively (Equation 3). Although, it is essential to remember that as the number of fixed pieces of evidence increases, the number of players who meet these criteria (i.e., probability of evidence) will decrease.

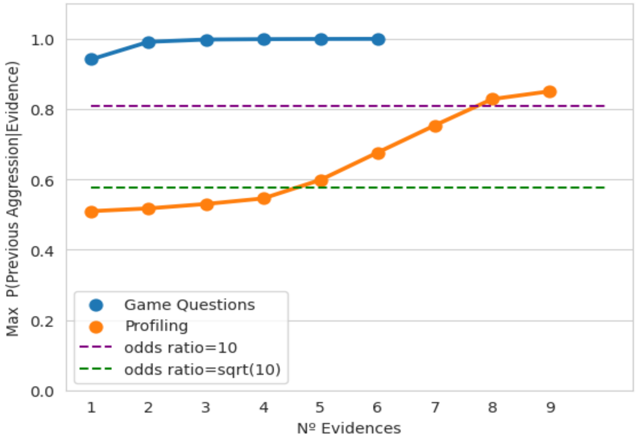


Figure 11. By performing a multi-factor analysis, we can find the combinations of variables that cause a greater increase in the conditional probability of the outcome (*Cyberthreat risk*). The following figure is obtained by finding the maximum conditional probability obtained by setting different numbers of combinations of evidence (from 1 to 9). This is done, on the one hand, for the variables obtained through the game questions and, on the other hand, for the profiling variables. The figure also shows the conditional probabilities corresponding to the relevant thresholds according to Jeffreys' criterion [14] calculated in Equation 3.

$$BF = \frac{\text{prior odds ratio}}{\text{posterior odds ratio}} = \frac{0.7/0.3}{(1-X)/X}$$
$$BF = 10^{1/2}(\text{ Substantial evidence }) \Rightarrow X \approx 0.57$$
$$BF = 10(\text{ Strong evidence }) \Rightarrow X \approx 0.81$$

Equation 3. Calculating the probability thresholds according to the Jeffreys criterion [14] and the selected prior

From 8 fixed pieces of evidence, the profile variables exceed the second threshold ~0.81. Further analysing this case, Figure 12 shows the number of observations of the risk profiles' most common shared profiling characteristics. We define a *risk profile* as one with a posterior probability (Cyberthreat risk = True) greater than or equal to 0.57. These risk profiles' top shared profiling characteristics are as follows: t*he profile that tends to share fake news is characterised by low honesty, male gender, high family support, medium neuroticism, low agreeableness, medium conscientiousness, and medium social support.*

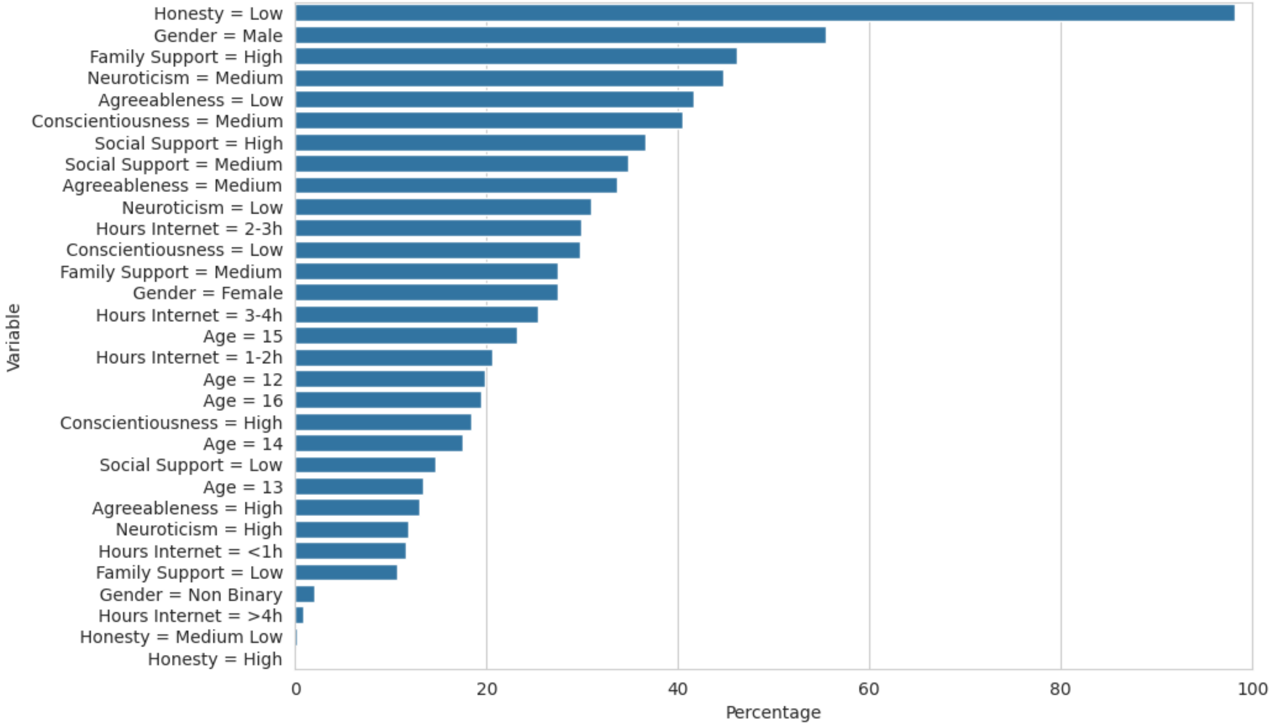


Figure 12. By performing a multi-factor analysis, we can find the combinations of variables (i.e., profiles) that the model has learned are "risky" for suffering cyberthreats. The figure shows a count of the number of times a particular value of a variable appears in the identified risk profiles. In this case, we analyse the profiles obtained by setting exactly 8 variables, since from this number onwards, we begin to obtain risky profiles with a relevant odds ratio according to Jeffreys' criterion [14].

5.4 Discussion

Based on the assumption that the proposed BN structures posited by experts in the field are accurate, the outcomes from the experiments provide compelling insights into the relationship between certain variables and patterns of behaviour related to cyberthreat victimisation risk.

Initially, the strength of each variable has been studied separately to explain the outcome of the variables of interest, i.e., cyberthreat victimisation risk. The study data clearly suggest that these video game-related questions have greater explanatory power compared to the profiling variables.

Secondly, the multifactorial analysis allows us to study the combined effects of different levels of each discrete variable, representing the various possible combinations of responses to the questions posed by the video game, as well as the combination of profile variable values. Our findings suggest that the questions posed are relevant and enable a more accurate discrimination with respect to the target variable compared to the other profile variables. The difference is pronounced, as evident in the graph, and with significantly fewer pieces of evidence, we achieve better segmentation.

However, in this case, unlike the CB case, there is no "ground truth" of the Cyberthreat victimisation risk variable. That is, now the variable of interest is latent. This implies that the methodology used can be interpreted as a Bayesian unsupervised clustering. That is, using the variables considered, we try to identify two distinct groups of people in the data. However, there is no guarantee that these groups actually correspond to minors at greater or lesser risk suffering the cybercrime under consideration. This also means that the values obtained in the experiments could be exaggerated or distorted.

6. Fake News

This section presents the methodology employed to analyse the prevalence of Fake News among adolescents using data from the RAYUELA video game sessions. As described in the Methodology section, BNs are utilised to model and estimate the causal relationships and probabilistic dependencies between the variables of interest. Fake news has emerged as a concerning phenomenon in the digital age, and its impact on adolescents is of particular concern. Fake news involves the dissemination of fabricated or misleading information presented as genuine news, often through social media platforms and other online channels. Given that adolescents are avid users of social media and the internet, they are increasingly exposed to a wide range of information, making them susceptible to the influence of fake news [22-23].

6.1 Data Processing

Data Source and Filters:

The data used in this analysis was collected from the RAYUELA's serious game pilots involving 726 participants. Within this adventure, adolescents encounter diverse scenarios related to fake news and respond to specific questions that shed light on their behaviours and decision-making when dealing with potentially misleading information. RAYUELA's serious game pilots have been designed to address the pressing issue of fake news and equip adolescents with media literacy skills to critically evaluate and discern reliable information from misinformation in the digital age. Through this engaging and interactive approach, we aim to gain a comprehensive understanding of how young users interact with fake news and develop insights to combat its influence effectively.

Gameplay Variables:

The selected variables primarily originate from Adventure 4, with some contributions from Adventures 2 and 3, within the Fake News analysis

- [Adventure 6] Fake News Check: In this first step, adolescents critically evaluate the credibility and accuracy of the webpage and its content by assessing its professionalism, source reputation, and data authenticity, and conduct additional research on the internet to corroborate the information.
- [Adventure 6] Web Page Looks Like: In this step, participants evaluate the professionalism and credibility of the webpage and information source to determine its authenticity.
- [Adventure 6] The Source: participants assess the reputation and trustworthiness of the news source to make judgments about its reliability.
- [Adventure 6] Information Looks Accurate: participants evaluate the accuracy of the presented information, considering the presence of substantial data and graphs, while also being cautious about the potential for falsification despite the visual representation.
- [Adventure 6] Replay Post: participants encounter a scenario where they decide on their response to provocative content, with some choosing not to engage and others proposing the addition of an anti-hoaxes website link to address the misinformation.
- [Adventure 6] Regarding Charles: participants either perceive it as challenging with limited options for resolution or propose attempting to talk to Charles as a way to address the issue.

Profiling (Socio-Demographic and Psychological) Variables:

In addition, our analysis includes socio-demographic factors like age, gender, migratory background, daily internet usage, social and family support, and psychological traits such as agreeableness, neuroticism, conscientiousness, and honesty. These variables help us understand adolescents' susceptibility to fake news and their interactions with misinformation, contributing to a comprehensive comprehension of this digital phenomenon.

6.2 DAG Design

By connecting all these variables as children of the 'FN_risk' variable, the DAG effectively models how the participants' perceptions and decision-making influence their overall risk of encountering fake news. The expert knowledge integrated into the video game's design is instrumental in capturing the complexities of fake news encounters, resulting in meaningful and relevant data that contributes to the analysis and understanding of fake news in the context of adolescent behaviour.

Figure 13 shows the BN architecture proposed by the experts for this case study, which is to analyse the variable of interest ‘**Fake News Sharing Risk**’. As the number of nodes in the network increases, it becomes more difficult to visualise and interpret. Therefore, to facilitate this task, we have listed below the indicators/variables considered and to which ones they causally affect:

- Age affects:
  - Daily Hours of Internet
  - Fake News Sharing Risk
  - CB Offending Risk
- Gender affects:
  - Fake News Sharing Risk
  - CB Offending Risk
- Migratory Background affects:
  - Fake News Sharing Risk
- Daily Hours of Internet affects:
  - Fake News Sharing Risk
  - CB Offending Risk
- Social Support affects:
  - Daily Hours of Internet
  - Fake News Sharing Risk
  - CB Offending Risk
- Family Support affects:
  - Daily Hours of Internet
  - Fake News Sharing Risk
  - CB Offending Risk
- Agreeableness (BF) affects:
  - Fake News Sharing Risk
- Neuroticism (BF) affects:
  - Fake News Sharing Risk
- Conscientiousness (BF) affects:
  - Fake News Sharing Risk
- CB Offending Risk affects:

    - Fake News Sharing Risk
- "Honesty" Question affects:
    - Game Answers
- Fake News Sharing Risk affects:
    - Game Answers

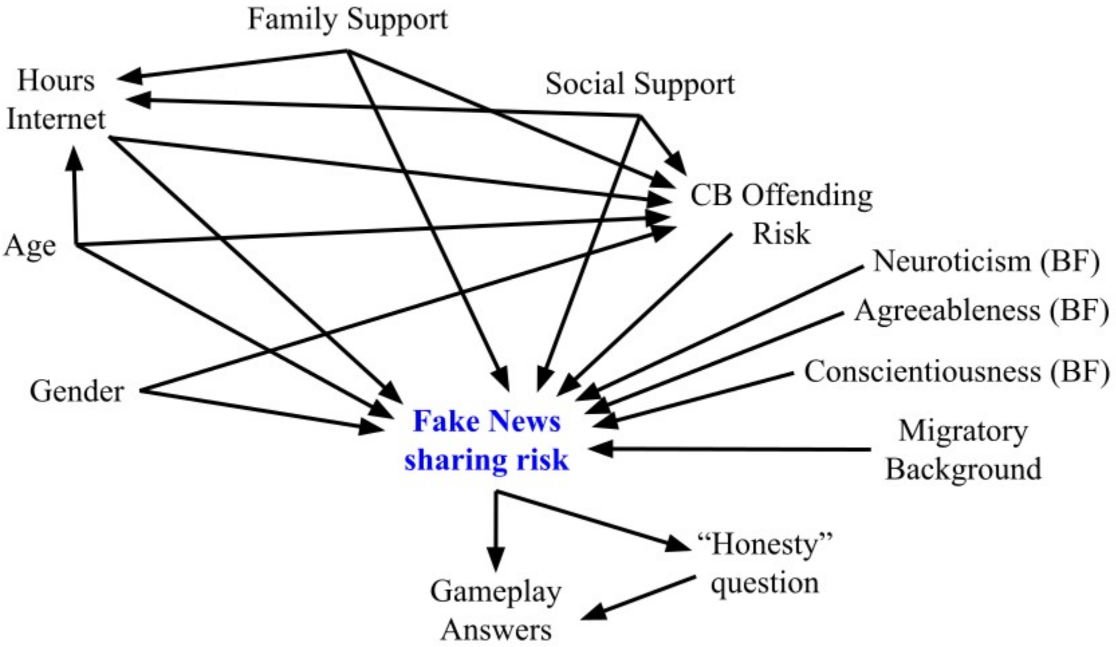


Figure 13. Structure of the Bayesian Network proposed by the experts of the RAYUELA project to analyse the variable of interest 'Fake News sharing risk'. The variable of interest in this case study is highlighted in blue.

6.3 Experiment Results

After formulating the most suitable causal framework for this specific case study and the variables involved, we move forward with conducting quantitative experiments based on causality principles.

6.3.1 Arrow Strength Analysis

As explained in the Methodology section, the influence analysis involves generating evidence in the Bayesian Network for the "Fake News risk" variable and comparing the probability distributions of the relevant indicators. Table 5 presents the outcomes of this analysis, utilising the Jensen-Shannon divergence to measure the impact of "Fake News risk" on the selected Bayesian Network structure designed by experts. The table displays normalised values, where 1 signifies full influence and 0 indicates no influence. While there is no standardised threshold for statistically significant influence, we have highlighted in green with an asterisk (*) those variables with values greater than 0.1, indicating their considerable influence.

Table 5. Influence Strength Analysis of Fake News Risk on the Expert-Proposed Bayesian Network Structure. The values are normalised, with 1 representing full influence and 0 indicating no influence. While there is no standard threshold for statistically significant influence, we have marked variables with values greater than 0.1 in green with an asterisk (*), signifying their notable impact. (See Annex III for more details on the content of the questions)

Indicators/Variable	Normalised Jensen-Shannon Distance
Adventure 6 Question 3: Source	0.42*
Adventure 6 Question 4: Information looks accurate	0.39*
Adventure 6 Question 2: Web page looks like	0.37*
Adventure 6 Question 6: Regarding Charles	0.30*
Adventure 6 Question 1: Migrant news check	0.21*
“Honesty” question	0.18*
Adventure 6 Question 5: Replay Post	0.15*
Conscientiousness (BF)	0.00
Neuroticism (BF)	0.00
Gender	0.00
Age	0.00
CB Offending	0.00
Daily Hours Internet	0.00
Agreeableness (BF)	0.00

6.3.2 Multi-Factor Marginalisation Analysis

As explained in the Methodology section, this analysis delves into the significance of variables in the model. However, unlike Arrow Strength Analysis, our focus here is on identifying multi-factor combinations that substantially alter the conditional probability of the variable 'Fake News risk.' Additionally, we aim to compare the relevance of variables derived from gameplay with those derived from demographic and psychological questionnaires.

Figure 14 illustrates the maximum conditional probability of a positive response to the 'Fake News risk' variable based on the quantity of evidence incorporated into the BN. The results are presented for both data sources: gameplay questions and profiling data. The graph includes two lines representing significant values of the conditional probability concerning the prior probability distribution, which was initially set at 0.3 before BN parameter training, with an effective sample size of 2 (soft prior).

Using the Jeffreys scale [14] for comparing odds ratios, a ratio between $10^{1/2}$ and 10 is interpreted as a substantial difference, while a ratio between 10 and $10^{3/2}$ indicates a strong difference. For our case, with a prior probability of 0.3, substantial differences occur with posterior probabilities of approximately 0.57, and strong differences occur with posterior probabilities of around 0.81 (Equation 3).

However, it is crucial to consider that as the number of fixed pieces of evidence increases, the number of players meeting these criteria (i.e., probability of evidence) will decrease.

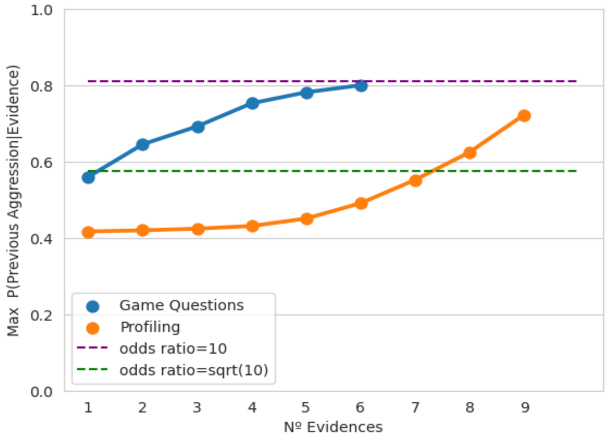


Figure 14. Through conducting a multi-factor analysis, we can identify the combinations of variables that lead to a significant increase in the conditional probability of the outcome (Fake News risk). The graph illustrates the maximum conditional probability achieved by considering varying numbers of evidence combinations (from 1 to 9). This analysis is performed for both the variables obtained from the game questions and the profiling variables. Additionally, the figure displays the conditional probabilities corresponding to the critical thresholds based on Jeffreys' criterion [14], as calculated in Equation 4.

$$
\begin{aligned}
& BF=\frac{\text{prior odds ratio}}{\text{posterior odds ratio}}=\frac{0.7/0.3}{(1-X)/X} \\
& BF=10^{1/2}(\text{ Substantial evidence }) \Rightarrow X \approx 0.57 \\
& BF=10(\text{ Strong evidence }) \Rightarrow X \approx 0.81
\end{aligned}
$$

Equation 4. Calculating the probability thresholds according to the Jeffreys criterion [14] and the selected prior

From the information provided, it appears that in the context of Fake News, the maximum probability is not achieved, primarily due to the limited number of questions from the adventure used, which hinders the identification of the underlying pattern. Nevertheless, by incorporating additional questions, particularly up to 7, it becomes plausible to attain the target threshold of approximately 0.81. In contrast, when analysing the profiling variables, we notice that the minimum level is reached with 7 variables, and the trend is still ascending. It is possible to reach the desired threshold of 0.81 by introducing a few more variables and evidence into the analysis, ultimately uncovering the profile with the highest risk of engaging in the dissemination of fake news is characterised by individuals with medium to low levels of honesty, no migratory background, medium neuroticism, male gender, low agreeableness, medium conscientiousness, high social support, and a history of CB offending

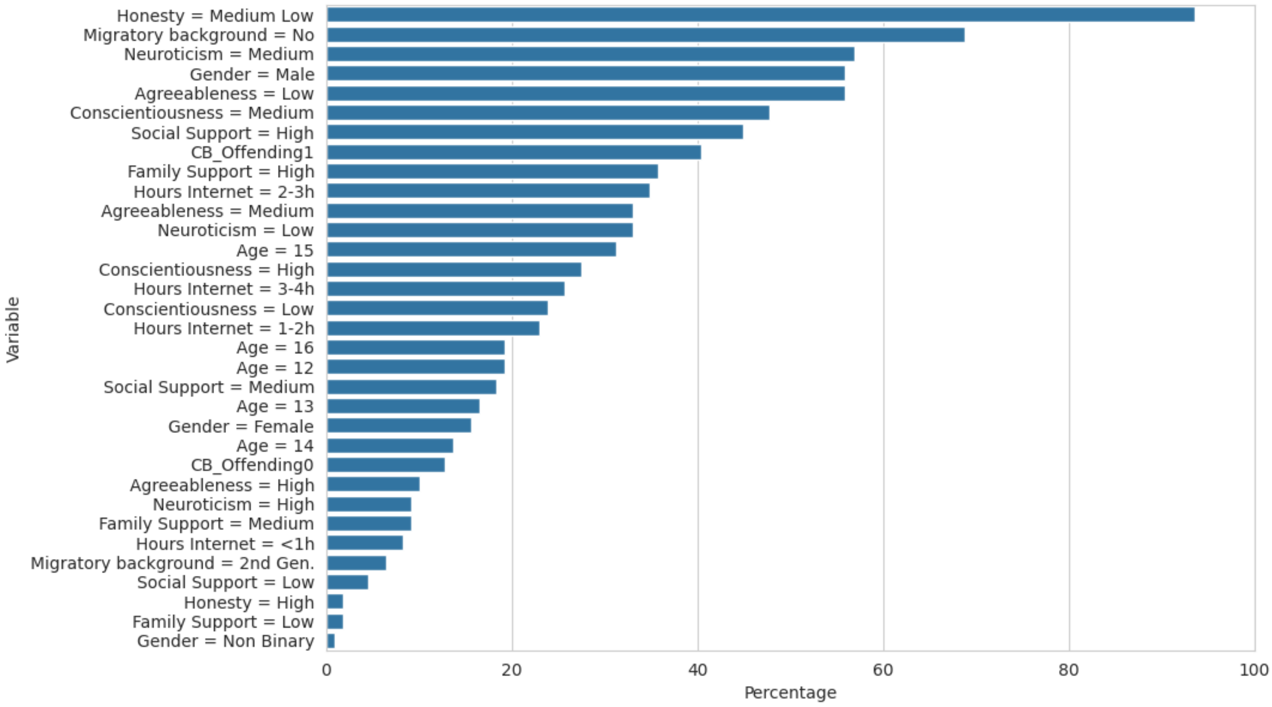


Figure 15. Through conducting a multi-factor analysis, we can identify the combinations of variables (i.e., profiles) that the model has recognized as "risky" for encountering fake news. The figure displays the frequency count of specific variable values found in the identified risk profiles. In this instance, we analyse the profiles achieved by precisely 11 variables, as from this point onward, we start to observe risky profiles with a substantial odds ratio as per Jeffreys' criterion [14]."

6.4 Discussion

The results obtained from the analysis provide valuable insights into the risk of encountering fake news among adolescents and shed light on the factors that influence their perception and decision-making. The DAG design proved effective in capturing the interconnections between variables, where all the expert-designed variables are connected as children of the 'FN_risk' variable. This design allows us to understand how the participants' assessments of news credibility and responses to provocative content collectively contribute to the overall risk of encountering fake news.

The expert-designed game questions, such as 'Adventure 6 Question 1: migrant news check,' ' Adventure 6 Question 2: web page looks like,' and ' Adventure 6 Question 3: the source,' demonstrate notable importance in the importance analysis. Adolescents' critical thinking and evaluation of news sources play a significant role in shaping their susceptibility to fake news. The consideration of factors like webpage professionalism and the reputation of the news source influences their trustworthiness judgments, which directly impact the 'FN_risk.'

Furthermore, variables like ' Adventure 6 Question 4: information looks accurate,' ' Adventure 6 Question 5: replay post,' and ' Adventure 6 Question 6: regarding Charles' reveal valuable information regarding participants' digital literacy and coping strategies. Adolescents who are cautious about assuming information accuracy and who choose not to engage with provocative content may display higher levels of media literacy, potentially lowering their overall risk of falling victim to fake news.

The findings provide valuable information for developing educational initiatives and interventions that aim to enhance adolescents' media literacy and equip them with the necessary skills to navigate the digital landscape responsibly. By empowering adolescents with knowledge and tools to identify and address fake news, we can foster a generation of informed digital citizens who are better prepared to engage with online information critically.

However, in this case, unlike the CB case, there is no "ground truth" of the fake news victimisation risk variable. That is, now the variable of interest is latent. This implies that the methodology used can be interpreted as a Bayesian unsupervised clustering. That is, using the variables considered, we try to identify two distinct groups of people in the data. However, there is no guarantee that these groups actually correspond to minors at greater or lesser risk suffering the cybercrime under consideration. This also means that the values obtained in the experiments could be exaggerated or distorted.

7. Discussion and Conclusions

7.1 Results summary

[Cyberbullying]

- **Research Question 1:** Which variables are most strongly related to the risk of suffering/committing cyberbullying?
  Based on the selected causal structure, available data, and metrics, the most relevant indicators of having **committed CB** are:
  - *Adventure 3 Question 3: Time Overrun*
  - *Adventure 3 Question 4: Pol Bullied*
  - *Adventure 3 Question 5: Remind Matthew*
  - *Adventure 1 Question 3: Matthew Meme*
  - *Previous Victimisation*
  - *Adventure 3 Question 1: Pirated Content*
  - *Adventure 3 Question 7: Help Pol*
  - *Adventure 3 Question 2: Pol Pola*

  Based on the selected causal structure, available data, and metrics, the most relevant indicators of having **suffered CB** are:
  - *Adventure 3 Question 5: Remind Matthew*
  - *Adventure 3 Question 3: Time Overrun*
  - *Adventure 3 Question 4: Pol Bullied*
- **Research Question 2:** What combinations of variables make it possible to construct meaningful risk profiles for suffering/committing cyberbullying?
  Based on the selected causal structure, available data, and metrics, the most relevant characteristics shared by risk profiles of having **committed CB** are:
  - *Previous CB victimisation=True*
  - *Gender=Male*
  - *Social Support = High*
  - *Family Support = High*

  Based on the selected causal structure, available data, and metrics, the most relevant characteristics shared by risk profiles of having **suffered CB** are:
  - *Sexual Orientation = Heterosexual*
  - *Migratory Background = No*
  - *Gender = Male*
  - *Social Support = High.*

[Online Grooming]

- **Research Question 1:** Which variables are most strongly related to the risk of suffering online grooming?
  Based on the selected causal structure, available data, and metrics, the most relevant indicators of being at **risk of suffering from OG** are:
  - *Adventure 2 Question 4: Place*
  - *Adventure 2 Question 8: Friend Request*
  - *Adventure 2 Question 9: Photos*
  - *Adventure 2 Question 3: Professional type*
  - *Adventure 2 Question 5: Profile photo*
  - *Adventure 5 Question 1: Secret*
  - *Adventure 2 Question 15: Block profile*
  - *Adventure 2 Question 1: Registration Name*
  - *Adventure 5 Question 2: Biology paper*
  - *Adventure 2 Question 14: Tell parents*
  - *Adventure 2 Question 12: Ask Help*
  - *Adventure 2 Question 11: More & more*
  - *Adventure 2 Question 7: Use PC*
- **Research Question 2:** What combinations of variables make it possible to construct meaningful risk profiles for suffering online grooming?
  Based on the selected causal structure, available data, and metrics, the most relevant characteristics shared by the profiles at **risk of suffering OG** are:
  - *Honesty = Low*
  - *Gender = Male*
  - *Family Support = High*
  - *Neuroticism = Medium*

[Cyberthreats]

- **Research Question 1:** Which variables are most strongly related to the risk of suffering online grooming?
  Based on the selected causal structure, available data, and metrics, the most relevant indicators of people at **risk of suffering cyberthreats** are:
  - *Adventure 4 Question 2: New Password*
  - *Adventure 2 Question 2: Registration Password*
  - *Adventure 4 Question 3: My Account Stolen*
  - *Adventure 4 Question 1: Phishing*
  - *Adventure 4 Question 4: Other Account Stolen*
  - *Adventure 3 Question 1: Pirated Content*
  - *"Honesty" Question*

- **Research Question 2:** What combinations of variables make it possible to construct meaningful risk profiles for suffering online grooming?
  Based on the selected causal structure, available data, and metrics, the most relevant characteristics shared by the profiles at **risk of suffering cyberthreats** are:
    - *"Honesty" = Low*
    - *Gender = male*
    - *Family support = High*
    - *Neuroticism = Medium*

[Fake News]

- **Research Question 1:** Which variables are most strongly related to the risk of suffering online grooming?
  Based on the selected causal structure, available data, and metrics, the most relevant indicators of people at **risk of sharing fake news** are:
    - *Adventure 6 Question 3: Source*
    - *Adventure 6 Question 4: Information Looks Accurate*
    - *Adventure 6 Question 2: Web Page Looks Like*
    - *Adventure 6 Question 6: Regarding Charles*
    - *Adventure 6 Question 1: Migrant News Check*
    - *"Honesty" Question*
    - *Adventure 6 Question 5: Replay Post*
- **Research Question 2:** What combinations of variables make it possible to construct meaningful risk profiles for suffering online grooming?
  Based on the selected causal structure, available data, and metrics, the most relevant characteristics shared by the profiles at **risk of sharing fake news** are:
    - *"Honesty" = Medium Low*
    - *Migratory Background = No*
    - *Neuroticism = Medium*
    - *Gender = Male*
    - *Agreeableness = Low*

7.2 Discussion and limitations

Task T6.3, in which this deliverable is framed, together with Task T6.2, were responsible for the quantitative analysis of the data collected through the RAYUELA pilots, so they are of great relevance to the project. Therefore, we believe that we should be cautious in drawing conclusions from the results. The sample size of the data is adequate for what is usual in social science research, but it is still relatively limited. Moreover, it should be kept in mind that questionnaire and video game data are often noisy and heterogeneous, especially when dealing with minors (e.g., some participants may have played randomly or deliberately answered questions incorrectly).

However, despite the limitations mentioned above, the results obtained seem to reveal interesting conclusions. In general, the results seem to show that the variables obtained through the video game are of great relevance in explaining and predicting the risk of suffering/committing the considered cybercrimes, thus, that the serious game has been designed correctly. Such video game variables are questions/situations/dilemmas designed by RAYUELA experts to capture relevant aspects regarding the considered cybercrimes.

At present, demographic indicators or variables obtained through registration do not appear to have significant relevance for either of the case studies. Except in the case of CB offending, where the variable indicating prior victimisation has a large influence in the experiments, both when the variables are considered individually and in combination (i.e., multi-factor). In other words, individuals who have previously been victims of CB exhibited a markedly higher propensity to commit CB offences. This relationship merits further study although it was one of the predictions of the work developed by WP1 to understand in depth the cybercrimes considered

As has been discussed repeatedly in this deliverable, it is critical to note that CB is the only cybercrime for which we have a "ground truth" [7] with which to validate the results obtained. For the rest of the cybercrimes, the methodology used is close to what could intuitively be described as Bayesian unsupervised clustering. That is, using the variables considered, we try to identify two distinct groups of people in the data. However, there is no guarantee that these groups actually correspond to minors at greater or lesser risk suffering the cybercrime under consideration. This also means that the values obtained in the experiments could be exaggerated or distorted.

By employing a (Bayesian) **causality-based methodology**, we can mitigate the biases present in the data correlations, assuming the selected BNs' structures to be accurate. This approach also forces us to make our assumptions and hypotheses explicit, leading to discussions and critical questions regarding the issues we are trying to address. This approach is especially important when dealing with sensitive research topics, such as those examined in RAYUELA. This approach differs from the Machine Learning-based approach used in Task T6.2. However, we see these approaches as complementary, helping us to better understand the problem from different perspectives.

7.3 Conclusions and future work

This deliverable presents an analysis of the data collected through the RAYUELA pilots using a **causality-based approach and Bayesian statistics**. The aim was to **identify key indicators/variables** that are relevant when discerning between potential victims and perpetrators of the considered cybercrimes.

It is critical to note that CB is the only cybercrime considered for which a validated psychological questionnaire is available for players to answer. This questionnaire [7] serves as a proxy for the underlying risk (the closest information that we can use as the “ground truth”). For the rest of cybercrimes, we have employed the exact same methodology, but without having their corresponding evaluation measure, so that the experiments could intuitively be described as Bayesian unsupervised clustering. That is, using the variables considered, we try to identify two distinct groups of people in the data. However, there is no guarantee that these groups actually correspond to minors at greater or lesser risk suffering the cybercrime under consideration. This also means that the values obtained in the experiments could be exaggerated or distorted.

We have received **assistance from experts** in the RAYUELA project to develop plausible BN structures for each cybercrime. Considering these structures as accurate, we performed a series of causal analyses based on the strength of influence, both looking at the individual influence of each variable and the combination of demographic and psychological variables to create risk profiles. This approach also forces us to make our assumptions and hypotheses explicit, leading to discussions and critical questions regarding the issues we are trying to address. This approach is especially important when dealing with sensitive research topics, such as those examined in RAYUELA.

Based on the findings, it appears that the **variables collected through the RAYUELA serious game are promising in detecting potential perpetrators and victims of the considered cybercrimes**. However, it is important to exercise caution in interpreting the results due to the limited amount of data available for analysis, as well as the potential noise inherent in social science and video game data. Nevertheless, these initial results suggest that the RAYUELA serious game has the potential to be a valuable tool for social research purposes, highlighting the need for further exploration of its capabilities.

Moving **forward**, several **research** areas should be explored to further enhance our understanding of the cybercrimes under consideration and the capabilities of the RAYUELA serious game. For example, it might be interesting to explore the simulation of interventions in Bayesian networks using J. Pearl's do-calculus [9]. In this way, we could rigorously simulate how possible interventions would affect minors to change their risk of suffering/committing cybercrime, moving from a predictive to a more prescriptive perspective.

The data collected through RAYUELA (appropriately pseudo-anonymized) has been made openly available (see D4.10) so that other researchers can use it and openly discuss the conclusions we have reached or apply different methodologies to it. All programming code used has been also made available to everyone through the GitHub of the project: https://github.com/rayuelaproject/Bayesian_Network_Analysis.

This work has mainly focused on the methodological development and the presentation of the results. The next logical steps are to discuss the possible practical implications and recommendations that can be given as a result of the obtained results. The results of such an exercise are reported in deliverable D6.6 and serve as input for the policy recommendations coming from this analysis reported in deliverable D7.7.

References

[1] Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan kaufmann.

[2] Pearl, J. (1995). Causal diagrams for empirical research. Biometrika, 82(4), 669-688.

[3] Russo, F. (2010). Causality and causal modelling in the social sciences. Dordrecht: Springer.

[4] Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of research in Personality, 41(1), 203-212.

[5] Rosenberg, M. (1965). Rosenberg self-esteem scale (RSE). Acceptance and commitment therapy. Measures package, 61(52), 18.

[6] Zimet, G. D., Dahlem, N. W., Zimet, S. G., & Farley, G. K. (1988). The multidimensional scale of perceived social support. Journal of personality assessment, 52(1), 30-41.

[7] Brighi, A., Ortega, R., Pyzalski, J., Scheithauer, H., Smith, P. K., Tsormpatzoudis, H., Tsorbatzoudis, H., et al. (2012). European Cyberbullying Intervention Project Questionnaire (ECIPQ) [Database record]. APA PsycTests.

[8] Glymour C, Zhang K and Spirtes P (2019) Review of Causal Discovery Methods Based on Graphical Models. Front. Genet. 10:524. Doi: 10.3389/fgene.2019.00524

[9] Pearl, J. (2012). The do-calculus revisited. arXiv preprint arXiv:1210.4852.

[10] Koiter, J. R. (2006). Visualizing inference in Bayesian networks. Delft University of Technology, 855.

[11] Endres, D. M., & Schindelin, J. E. (2003). A new metric for probability distributions. IEEE Transactions on Information theory, 49(7), 1858-1860.

[12] Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. The annals of mathematical statistics, 22(1), 79-86.

[13] Georgios N. Yannakakis, & Togelius, J. (2018). Artificial Intelligence and Games. Springer.

[14] Jeffreys, H. (1998). The theory of probability. OuP Oxford.

[15] Cai, B., Huang, L., & Xie, M. (2017). Bayesian networks in fault diagnosis. IEEE Transactions on industrial informatics, 13(5), 2227-2240.

[16] Marcot, B. G., & Penman, T. D. (2019). Advances in Bayesian network modelling: Integration of modelling technologies. Environmental modelling & software, 111, 386-393.

[17] Kabir, S., & Papadopoulos, Y. (2019). Applications of Bayesian networks and Petri nets in safety, reliability, and risk assessments: A review. Safety science, 115, 154-175.

[18] Arató, N., Zsidó, A. N., Rivnyák, A., Péley, B., & Lábadi, B. (2022). Risk and protective factors in cyberbullying: the role of family, social support and emotion regulation. *International journal of bullying prevention*, *4*(2), 160-173.

[19] Zhu, C., Huang, S., Evans, R., & Zhang, W. (2021). Cyberbullying among adolescents and children: a comprehensive review of the global situation, risk factors, and preventive measures. *Frontiers in public health*, *9*, 634909.

[20] Ioannou, A., Blackburn, J., Stringhini, G., De Cristofaro, E., Kourtellis, N., & Sirivianos, M. (2018). From risk factors to detection and intervention: a practical proposal for future work on cyberbullying. *Behaviour & Information Technology*, *37*(3), 258-266.

[21] Barkoukis, V., Lazuras, L., Ourda, D., & Tsorbatzoudis, H. (2016). Tackling psychosocial risk factors for adolescent cyberbullying: Evidence from a school-based intervention. *Aggressive behavior*, *42*(2), 114-122.

[22] Papapicco, C., Lamanna, I., & D'Errico, F. (2022). Adolescents' vulnerability to fake news and to racial hoaxes: a qualitative analysis on italian sample. *Multimodal Technologies and Interaction*, *6*(3), 20.

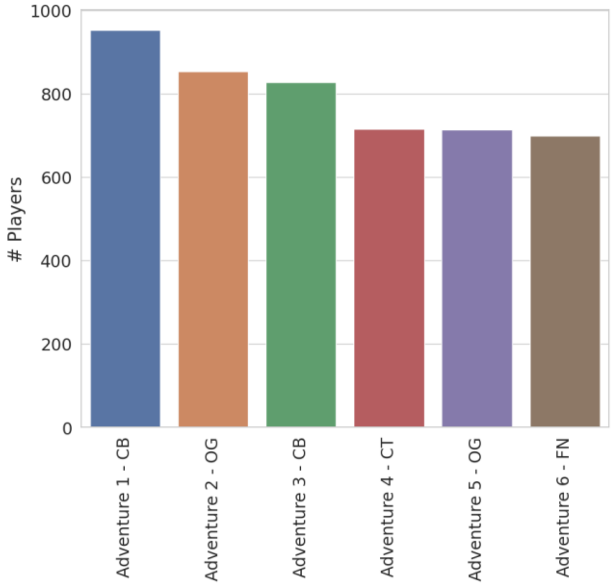
[23] Herrero-Diz, P., Conde-Jiménez, J., & Reyes-de-Cózar, S. (2021). Spanish adolescents and fake news: level of awareness and credibility of information (Los adolescentes españoles frente a las fake news: nivel de conciencia y credibilidad de la información). *Culture and Education*, *33*(1), 1-27.

Annex I: Exploratory Data Analysis (3rd pilot phase)

Below is a set of descriptive statistics on the data collected up to the third phase of the RAYUELA pilots. Gameplay data has not been considered in this exploratory analysis. We only considered demographic data and the psychological/sociological questionnaires that the students had to fill in before and after playing the game.

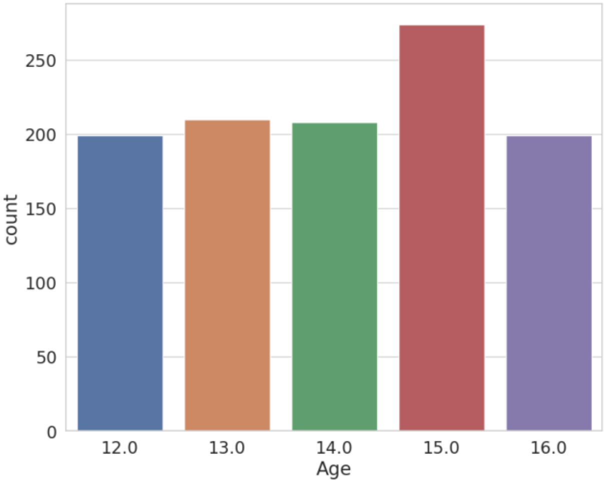
Number of players

Total Players in the registry: 1147		
Adventure	**# Players**	**Percentage**
Adventure 1 - CB	953	83.1%
Adventure 2 - OG	853	74.4%
Adventure 3 - CB	828	72.2%
Adventure 4 - CT	716	62.4%
Adventure 5 - OG	714	62.3%
Adventure 6 - FN	699	60.9%



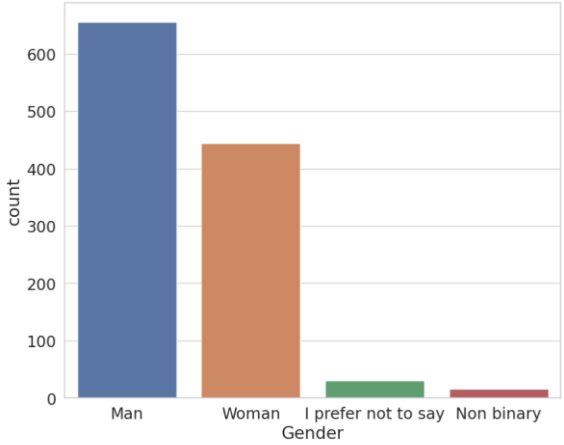
Age

Age	# Players	Percentage
12	199	17.35%
13	210	18.3%
14	208	18.1%
15	274	23.9%
16	199	17.35%



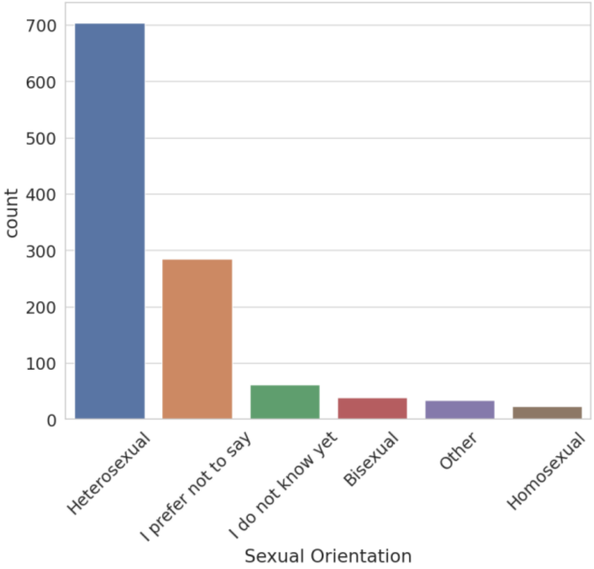
Gender

Gender	# Players	Percentage
Man	656	57.2%
Woman	444	38.7%
"I prefer not to say"	31	2.7%
Non-Binary	16	1.4%



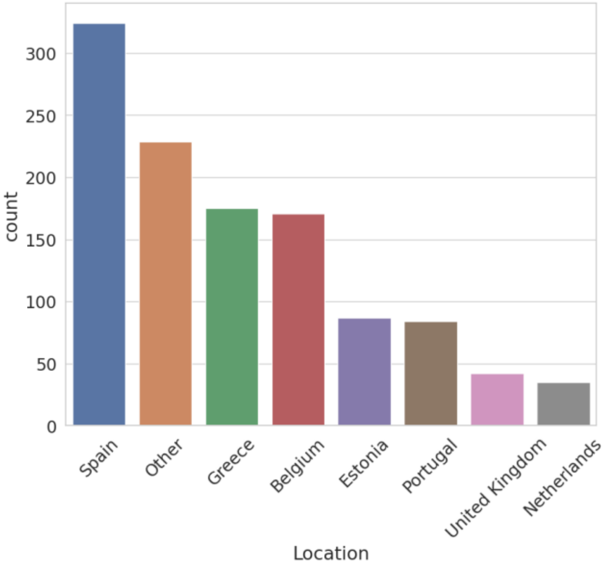
Sexual Orientation

Sexual Orientation	# Players	Percentage
Heterosexual	704	61.4%
"I prefer not to say" / Not asked	285	24.8%
"I don't know yet"	62	5.4%
Bisexual	39	3.4%
Other	34	3%
Homosexual	23	2%



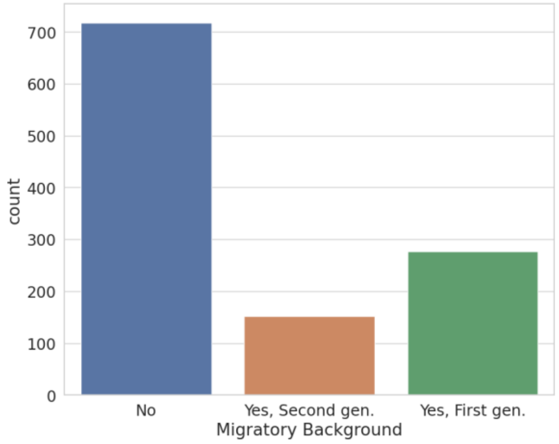
Country

Sexual Orientation	# Players	Percentage
Spain	324	28.25%
Other	229	20%
Greece	175	15.25%
Belgium	171	14.9%
Estonia	87	7.6%
Portugal	84	7.3%
United Kingdom	42	3.7%
Netherlands	35	3.05%



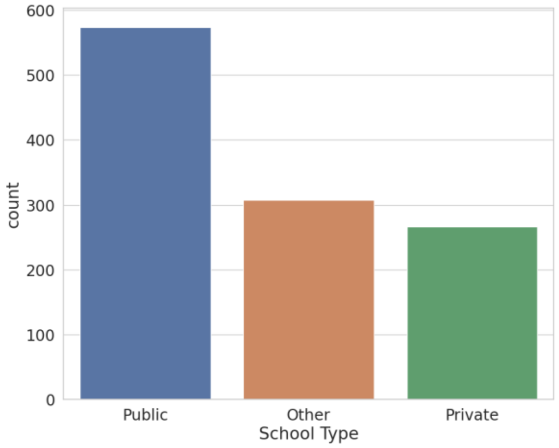
Migratory Background

Migratory Background	# Players	Percentage
No	718	62.6%
Yes, First Gen.	277	24.15%
Yes, Second Gen.	152	13.25%



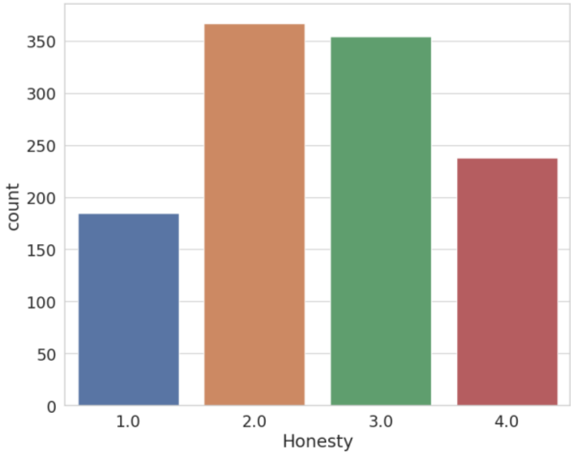
School Type

Migratory Background	# Players	Percentage
Public	574	50%
Other	307	26.8%
Private	266	23.2%



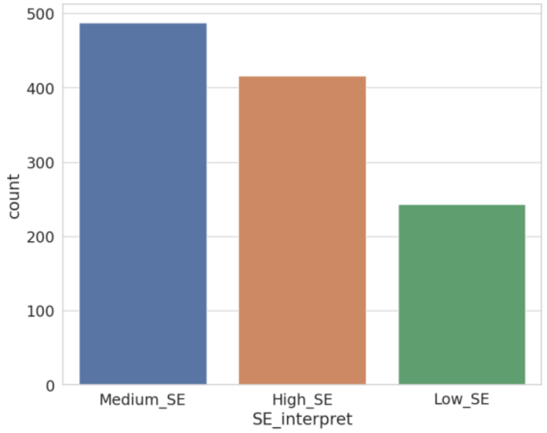
"Have you played like you would behave in real life?"

Have you played like you would behave in real life?	# Players	Percentage
1 – very different	185	16.2%
2 – different	367	32.1%
3 – similar	354	30.9%
4 – very similar	238	20.8%



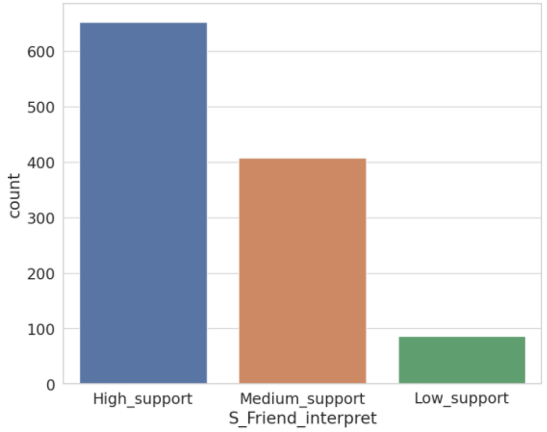
Self-Esteem

Self-Esteem	# Players	Percentage
Medium	488	42.5%
High	416	36.3%
Low	243	21.2%



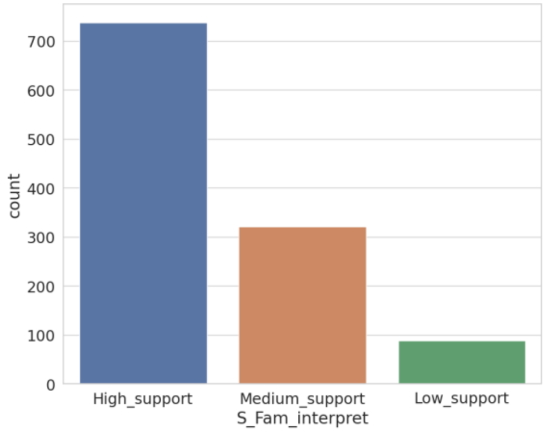
Social Support

Social Support (Friends)	# Players	Percentage
High	653	56.9%
Medium	408	35.6%
Low	86	7.5%

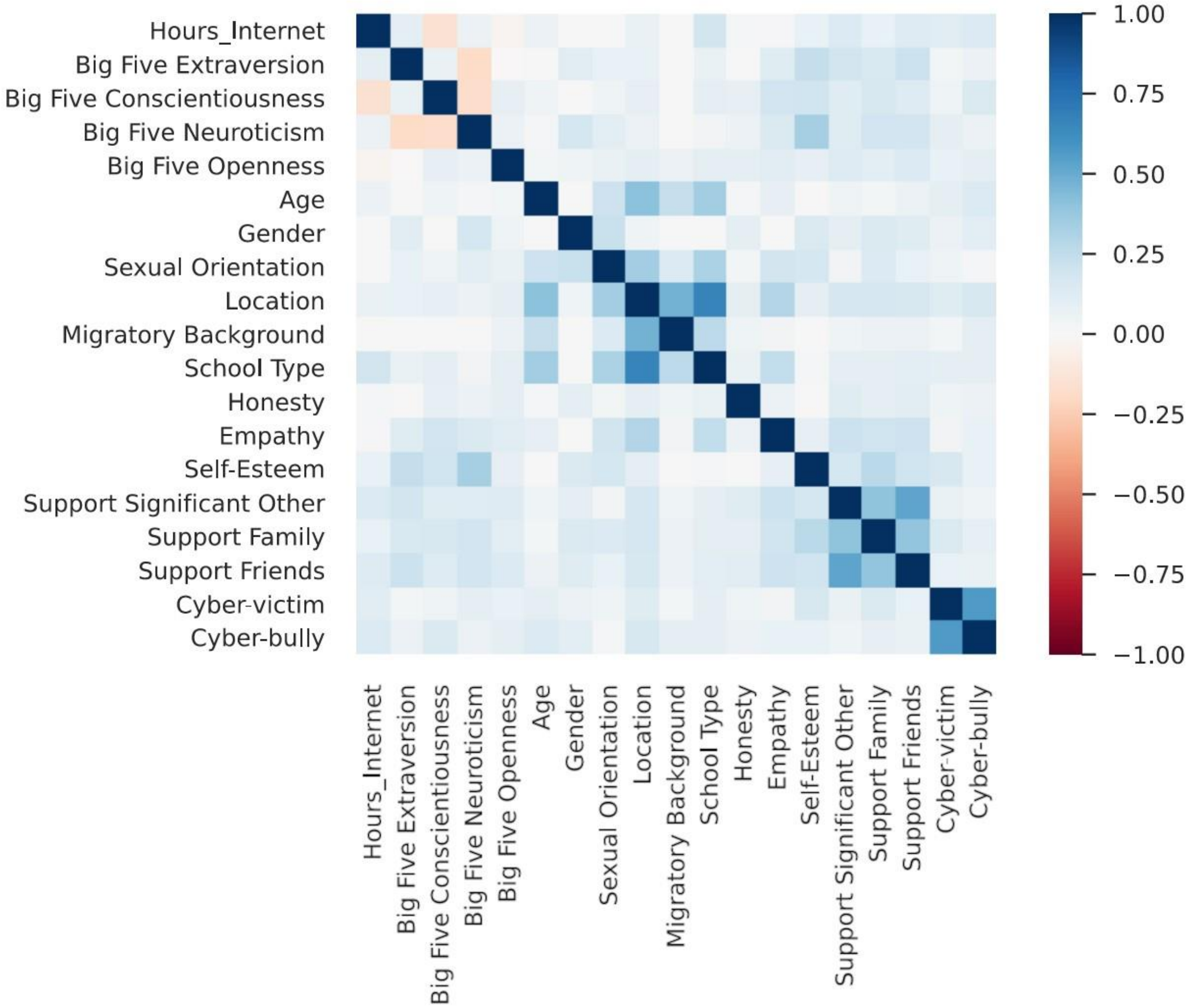


Family Support

Social Support (Friends)	# Players	Percentage
High	738	64.3%
Medium	321	28%
Low	88	7.7%



Correlations between variables



Significant correlations:

- "Location" is highly correlated with "School type"
- "Cyber-victim" is highly correlated with "Cyber-bully"
- "Support Friends" is highly correlated with "Support Significant Other"

Annex II: Exploration of Latent Variables and Validation Using t-SNE

In the data analysis phase of our project, we have employed the technique of t-SNE (t-Distributed Stochastic Neighbour Embedding) to explore and validate our initial hypotheses. Our goal has been to apply EM learning to a Bayesian network for identifying latent variables that describe cybercrimes based on responses in a video game and socio-demographic data.

Functionality of t-SNE:

t-SNE is a nonlinear dimensionality reduction algorithm that has allowed us to visualize complex data in a two-dimensional space. The algorithm aims to preserve similarity relationships between points in the original space and the reduced space by minimizing the Kullback-Leibler divergence between associated probability distributions. This aids us in revealing clusters and patterns in our data.

Validation and Data Exploration:

We have employed t-SNE as a descriptive and exploratory tool to validate the presence of clusters in our data based on latent variables from the Directed Acyclic Graph (DAG). Categorical variables were transformed into numerical representations before applying t-SNE, recognizing that this algorithm may not capture all subtle relationships between such variables.

Utilizing t-SNE in Validation:

One of the ways t-SNE assists us in validating the existence and diverse identifications of latent variables is through the resulting visualisations. Clusters identified in these visualisations suggest that specific combinations of game responses and socio-demographic data lead to the formation of distinct clusters in the reduced space. These clusters represent different values of the latent variables we seek.

For instance, in our Fake News study, t-SNE revealed distinct clusters based on responses to certain questions, supporting the existence of different player profiles related to this type of cybercrime. This indirectly validates the presence of underlying latent variables influencing player decision-making, which are captured by our observable variables.

In another example, within the context of Online Grooming, t-SNE revealed a small group of users who had responded to a limited number of questions. However, we did not find a variable that clearly distinguishes two population groups.

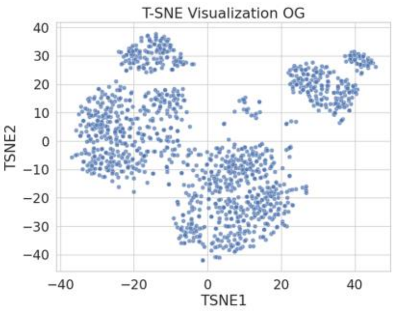
In summary, the integration of t-SNE into our methodology has proven to be an effective strategy for exploring and validating the presence of clusters in our data. The combination of t-SNE with our most relevant variables has allowed us to assess coherence between latent and observable structures. While results should be interpreted with caution, t-SNE has significantly contributed to our understanding of relationships within our data.

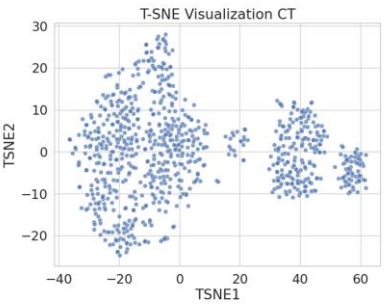
Results for Each Analysed Cybercrime:

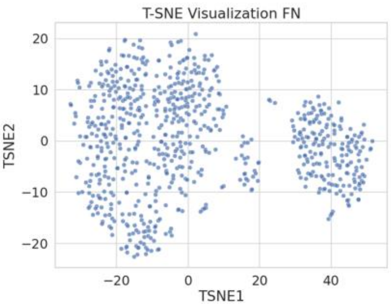
For each of the analysed cybercrimes, we have constructed four graphical representations based on the data used for training and generating network parameters (prior to discretization). Utilizing the necessary pre-processing step of one-hot encoding, we have generated unique t-SNE visualizations tailored to each specific cybercrime.

The outcomes have yielded intriguing insights, as evident in Graphs 1, 2, and 3, corresponding to the two-dimensional plots for Fake News, Online Grooming, and Cyber Threats, respectively, where the two corresponding coordinates are represented.

Of particular interest is the observation that distinct groups can indeed be identified within each case. This initial validation of separable groups (in topological terms) is crucial to ensure that the structure of conditional independence defined around latent variables remains meaningful.





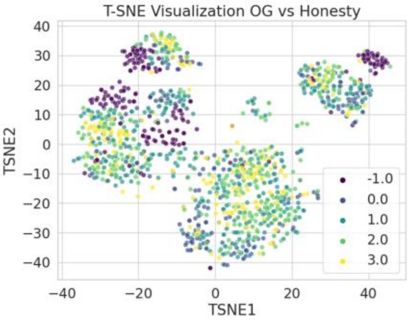


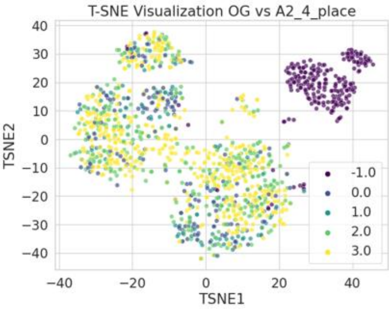
Interestingly, we find that there are more groups than levels within these latent categorical variables. This suggests that these groups likely do not align directly with the levels of latent variables constructed across various networks.

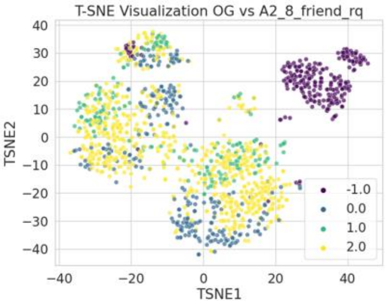
To further explore the relationship, we employ colour-coded points in these plots to represent the four most significant variables for each cybercrime. These variables have been identified as crucial in distinguishing between different cybercrimes. Below, we provide commentary on the conclusions drawn from each case:

Variables for Each Cybercrime:

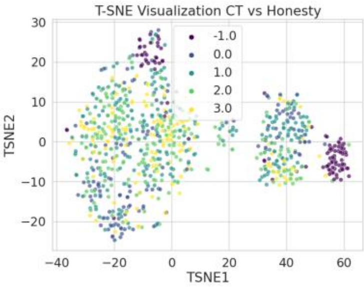
Online Grooming (OG): The variables consist of "Honesty," "Adventure 2," "C4 registration place," and "Adventure 2 friend request." In this case, the visualisation reveals a subgroup of users who responded to a limited number of questions. However, unlike the cases of Fake News and Cyber Threats, we did not find a variable that clearly differentiates two population groups.

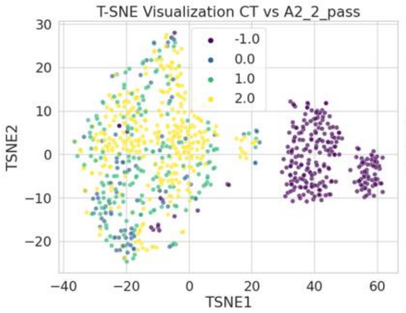


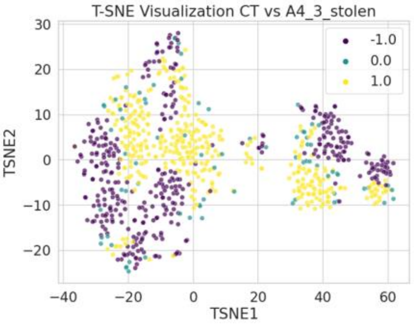




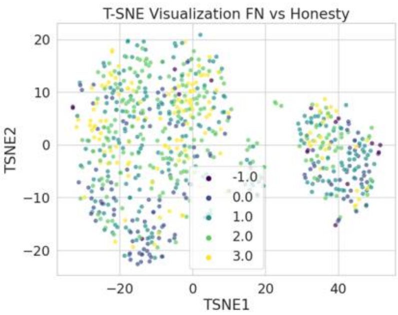
Cyber Threats (CT): The variables are "Honesty," "Adventure 2," "C2 2 Registration password," "Adventure 4 new password," and "Adventure 4 Account stolen." Notably, responses indicating non-completion of the password renewal process seem relevant, indicating their membership in one group. Additionally, regarding account theft, there appear to be two possibilities: experiencing theft or not responding, thus suggesting the identification of distinct groups. Given that this question is closely related to the latent variable for this cybercrime, these results imply the potential identification of the variable.

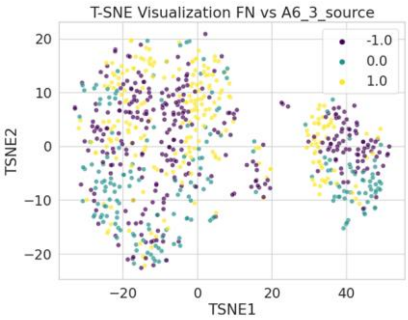


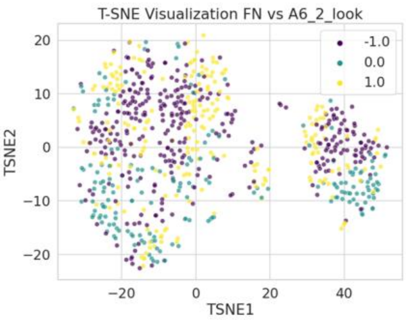




Fake News (FN): The variables include "Honesty," "Adventure 6 the source," "Adventure 6 web page look like," and "Adventure 6 information look accurate." While the variable related to information does not distinctly differentiate groups, the ones concerning the appearance of the source seem to have distinguishable classes, employing one of the t-SNE components.







Annex III: Game decisions transcript

Adventure 1 - CB

Question 1: Photo Sharing

[Talking to Matthew after taking a selfie.]

Now we only have to share and tag the photos. Jane, do you want to share them, or do you prefer me to do it?

□ *I will do it.*

□ *You can do it.*

Question 2: Sociable

[Talking to Robert after sharing the selfie. Dialog depends on the previous answers.]

It seems you like to upload many photos and share stuff on your social network.

□ *I would say I am sociable.*

□ *I consider myself kind of shy.*

Question 3: Matthew Meme

[After receiving a message from Patty with the meme about Matthew.]

□ *Hehe, it's funny, I will share the meme.*

□ *I am not going to share it.*

□ *I won't share the meme and I'm going to try to end up with this.*

Adventure 2 - OG

Question 1: Registration Name

[Creating a new profile on a social network. The user has to select a profile name.]

□ *My name and my year of birth.*

□ *My name and surname.*

□ *My favourite music band name.*

□ *Other famous/TV/Book character I like.*

Question 2: Registration Password

[The user has to select a profile password.]

□ *I don’t have time for this; I will leave the default password.*

□ *My name and surname.*

□ *I’ll use the same password I have on other websites, so it’s easier to remember.*

□ *I am going to set a strong password, even if I have to invest some more time.*

Question 3: Registration Profile Type

[The user has to select a profile type.]

□ *Public profile.*

□ *Private profile.*

Question 4: Registration Profile Place

[The user has to select a profile place.]

□ *The name of the city and neighbourhood where I live.*

□ *The name of my school and country.*

□ *Something fantastic, as "I am in the clouds", "In the moon" or "Too far away from X".*

□ *Leave empty.*

Question 5: Registration Profile Photo

[The user has to select a profile photo.]

□ *A photo of just me.*

□ *A photo of me and some friends.*

□ *A photo from the Internet, in which I do not appear.*

Question 6: Comment Patty Post

[Seeing a post from Patty on the social network.]

□ *Good one!*

□ *They are awesome.*

□ *I don’t like them, it’s so childish.*

□ *Don’t send any comment. I’m sure she won’t pay any attention to it.*

Question 7: Use PC

[Using the club’s PC after accepting a friend request from a photographer on the social network.]

□ *View messages.*

□ *Check photographer’s profile.*

Question 8: Friend Request

[Using the club's PC.]

□ *Accept friend request.*

□ *Reject friend request.*

□ *Check photographer's profile.*

Question 9: Send Photos

[Checking messages after accepting the friend request.]

□ *Send photos.*

□ *Not send photos.*

Question 10: More Photos

[Checking messages after sending swimsuit photos.]

□ *Send naked photos.*

□ *Do not send naked photos.*

Question 11: More & More

[Checking messages after sending naked photos.]

□ *Send more naked photos.*

□ *Reject the request and inform Mary.*

Question 12: Ask Help

□ *Ask for help to Mary.*

□ *Say nothing.*

Question 13: Close Case

□ *Ok, I'll check the profile.*

□ *No, I won't check the profile.*

Question 14: Tell Parents

[At the end of the scene. Previous dialog depends on player decisions.]

We should report that profile to the social network. Besides, you should also tell your parents about it and see if we need to talk with the police.

□ *I don't know, Mary, communication with them is not very easy. Lately they get angry about anything and we always end up shouting.*

□ *I'm ashamed! I don't want to tell them something like that. It's better if I try to solve it on my own.*

□ *Yeah, you're probably right. I'm a bit embarrassed but I'll give it a try.*

Question 15: Block Profile

□ *Block the profile.*

□ *Do not block the profile.*

Adventure 3 – CB

Question 1: Pirated Content

[Playing video games in your room.]

I know of some sites that pirate the content and then you can download the update for free.

□ *I will download the pirated update through a website.*

□ *I will wait until I have money or until my parents give me the money to buy the new expansion.*

Question 2: Pol or Pola

[Playing video games in your room. Your friends joke about Pol's appearance.]

□ *I don't like this kind of jokes.*

□ *That's funny.*

□ *Say nothing.*

Question 3: Time Overrun

[Playing video games in your room. A warning pops up about the number of hours you have been online]

□ 4 hours are not that much. So, I can keep chatting a bit longer.

□ It's time to stop and disconnect for a while, although I might miss some juicy gossiping.

Question 4: Pol Bullied

I heard that some guys are messing with Paul. I am worried that he may be bullied. What do you think?

□ *They are just having fun; I would not call that bullying.*

□ *I think it's not right... but calling that bullying is a bit of a stretch.*

□ *I think that's unacceptable; we should do something about it.*

Question 5: Remind Matthew

[Your friends start messing with Pol.]

□ *Yes, I had a similar bad experience... I don't like being picked on.*

□ *No, it has never really happened to me, to my knowledge.*

□ *Yes, it was me who messed with someone else... but it was not such a big deal.*

Question 6: Talk to Pol

So, shall we talk to Pol to see how he is?

□ *It is better to let him be.*

□ *Of course, we should try to help.*

Question 7: How to Help Pol

We can help you. You are not alone in this. I think...

□ *We should go to tell the teacher, he should know what to do.*

□ *We should report the comments to the social network, so that it doesn't happen again.*

□ *We should not report it, because I don't want to get picked on for being a snitch. . .*

□ *We should not report it as reporting is usually useless.*

Adventure 4 -CT

Question 1: Phishing Email

[You receive an email indicating that your social network account has been compromised.]

□ *Is my account in danger?! I must act quickly before I lose it. I will follow the instructions in the email.*

□ *I find it suspicious...I'll better go straight to my social network profile's security settings and change my password there.*

Question 2: New Password

[You proceed to change the password of your account.]

□ *(Password = Name123) I'm going to leave a password very similar to the one I had before. Otherwise, I'll forget it...*

□ *(Password = Football10) I'm going to make a password with some hobbies or things I like in it. So, I won't forget it!*

□ *(Password = Ax/2oP3%nY6) I'm going to make my password difficult and long. It is more challenging this way, but much safer.*

Question 3: My Account Stolen

[If your account has been phished.]

□ *It does not seem to be that worrying, it's just a social network account. We don't need to tell this to anyone. I can create another account after all.*

□ *It is important to tell someone or report it, since the account contains personal information. It is a crime!*

Question 4: Other Account Stolen

[If John has not been phished.]

□ *It does not seem to be that worrying, it's just a social network account. We don't need to tell this to anyone. He can create another account after all.*

□ *It is important to tell someone or report it, since his account contains personal information. It is a crime!*

Adventure 5 – OG

Question 1: Secret Relationship

[You and your friends are commenting that Sheila has a new romantic relationship that is distancing her from her friends and you are worried.]

□ *Love is love, and everyone experiences it in a different way. If she needed help, she would have asked for it, wouldn't she?*

□ *Sounds a bit creepy to me, have you tried looking at her social media?*

Question 2: Biology Paper

[You must meet to do a biology assignment and indicate your preference to meet online or in person.]

□ *Meet at the library this afternoon, so we can go to the cafeteria if we finish earlier.*

□ *Do it by video-conference this afternoon, so we can be more comfortable at home.*

Question 3: Talk Sheila

[John sees Sheila, who is leaning against the wall with her eyes fixed on her mobile phone.]

□ *I will talk to her and let her know she can trust me if she has any problem.*

□ *I should text Mary since she is closer to Sheila.*

Adventure 6 - FN

Question 1: Migrant News Check

[When investigating a news item that appears to be false, you must decide which things seem most relevant to verify the information.]

□ *How professional the web page looks like (style, images, design, etc).*

□ *The source itself: is it a known newspaper/website or is it an unknown site?*

□ *If the information looks accurate, for instance with enough numbers and statistics.*

□ *Search on the Internet to contrast the information.*

Question 2: Web Page Looks Like

[Reviewing the website.]

□ *It seems it is true. Definitely not a fake page.*

□ *It looks quite professional, but does it mean it's not fake? We should try other options.*

Question 3: The Source

[Reviewing the source.]

□ *It is a known newspaper, at least I've seen it quite a lot on Social Networks. I would say is not fake.*

□ *Even though it is a kind of famous newspaper, it could contain fake information. We should try other options.*

Question 4: Information Looks Accurate

[Reviewing if the information looks accurate.]

□ *Ok, they are displaying a big amount of data, and look at the graph as it rises. It looks pretty accurate.*

□ *Ok, there are a lot of numbers and graphs, but that does not mean that the data is correct. Data can also be falsified.*

Question 5: Replay post

Ok, so first of all, we should report the content to the social network, and we should probably reply with this information, right?

□ *It is not worth answering. Don't feed the troll!*

□ *Yes, let's add the link to the anti-hoaxes' website.*

Question 6: Regarding Charles

What should we do with Charles?

□ *It is a basket case, there is little we can do.*

□ *We should try to talk to him.*