Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

Mathias Øveren Enger

Ranking The Stars

Combining Graph Theory and Fuzzy Logic to Detect Cybergroomers

Master's thesis in Cybersercurity and Data Communcation
Supervisor: Patrick Bours
June 2024

NTNU
Norwegian University of
Science and Technology

Mathias Øveren Enger

Ranking The Stars

Combining Graph Theory and Fuzzy Logic to Detect Cybergroomers

Master's thesis in Cybersercurity and Data Communcation
Supervisor: Patrick Bours
June 2024

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

Ranking the "Stars": Combining Graph Theory and Fuzzy Logic to Detect Cybergroomers

Øveren Enger, Mathias

Submission date: June 2024
Main supervisor: Bours, Patrick, NTNU

Norwegian University of Science and Technology
Department of Information Security and Communication Technology

Title: Ranking the "Stars":
Combining Graph Theory and Fuzzy Logic to Detect Cybergroomers

Student: Øveren Enger, Mathias

Problem description:

”On the internet nobody knows you’re a dog” is the title of a famous cartoon by Peter Steiner, published in The New Yorker on the 5th of July 1993. Now, nearly 30 years later, there are a lot of ”dogs” present on the internet with less then good intentions. We are in particular looking at users in online chat platforms to detect sexual predatory conversations as early as possible. We do not evaluate conversations afterwards (a forensic approach), as then the children might have fallen victim to sexual abuse already.

Besides evaluating conversations are we also looking at evaluating the behaviour of users in online chat fora. An ordinary user will chat with a (limited) number of friends in relatively equal amounts (so similar amounts of messages to each chat partner), while the chatpartners of an ordinary user will often also be connected. A sexual predator will try to contact many different potential victims and will in most cases not get a reply. In some cases he/she might get a reply after which the child quickly determines that they are not interested in further communication. In rare cases the groomer will have a more extensive contact with the victim. This chat pattern is very skewed and thereby detectable.

The research so far has concentrated mostly on trying to detect single users in a chat and determining the risk that their behaviour is malicious or not. This current research project will actually look at the ”ranking” of the all the users on a platform. Particular behaviour will increase the risk that the user is malicious and hence this user should be ranked higher up, while other (innocent) behaviour will move the user lower on the ranking system. In such a real-life system the highest ranked users can then be evaluated by moderators, but in the research work the students should determine if malicious users will end up high on the ranking or not.

Approved on:	2024-02-12
Main supervisor:	Bours, Patrick, NTNU

Abstract

As children increasingly immerse themselves in the digital world, they face new risks different from those in the physical world. Unlike in the physical world, where people can see and verify the identity of the person they are speaking to, the digital world often allows individuals to disguise their true identities. This anonymity enables child predators to pose as minors, gaining the trust of unsuspecting young users and ultimately exploiting them sexually. The rise in such malicious activities underscores the urgent need for effective detection and prevention mechanisms, with early detection being crucial to intervene before harm occurs.

In this thesis, we have tackled the challenge of early cybergrooming detection by integrating graph theory with fuzzy logic to analyze user behavior in an action-based ranking system. We created a decision tree built on the principles of fuzzy logic, which serves as an analysis tool for every interaction a user is involved in. For every interaction, the user's risk score is updated, allowing for continuous assessment. This system allows us to dynamically rank users, with those exhibiting the most predatory behavior receiving the highest risk scores and being ranked near the top.

The result of this study is that continuous analysis of user behavior is effective in detecting unwanted users such as sexters, spammers, and sexual predators. Our system can rank the majority of these users among the top ranks, effectively highlighting those who exhibit the highest risk based on their interaction patterns. This system also opens up numerous opportunities for future research. Integrating conversation analysis with behavior analysis may further enhance the system, which could enable a more comprehensive understanding of user interactions.

Sammendrag

Etter hvert som barn blir stadig mer involvert i den digitale verden, står de overfor nye risikoer som er forskjellige fra de i den fysiske verden. I motsetning til virkeligheten, hvor man kan se og verifisere hvem man snakker med, tillater den digitale verden ofte enkeltpersoner å skjule sin sanne identitet. Denne anonymiteten gjør det mulig for overgripere å utgi seg for å være barn, få tillit fra intetanende unge brukere og til slutt utnytte dem seksuelt. Økningen i slike skadelige aktiviteter understreker det akutte behovet for effektive deteksjons- og forebyggingsmekanismer, der tidlig deteksjon er avgjørende for å kunne gripe inn før skade oppstår.

I denne oppgaven har vi tatt for oss utfordringen med tidlig deteksjon av cyber-grooming ved å integrere grafteori med fuzzy logikk for å analysere brukeradferd i et handlingsbasert rangeringssystem. Vi har laget et beslutningstre basert på prinsippene i fuzzy logikk, som fungerer som et analyseverktøy for hver interaksjon en bruker er involvert i. For hver interaksjon blir brukerens risikoscore oppdatert, noe som muliggjør kontinuerlig vurdering. Dette systemet lar oss dynamisk rangere brukere, der de som viser mest overgriper-lignende adferd får de høyeste risikoscorene og blir rangert nær toppen.

Resultatet av denne studien er at kontinuerlig analyse av brukeradferd er effektivt for å oppdage uønskede brukere som sextere, spammere og seksuelle rovdyr. Systemet vårt kan plassere flertallet av disse brukerne blant de øverste rangeringene, og effektivt fremheve de som viser høyest risiko basert på deres interaksjonsmønstre. Denne løsningen åpner også opp for mange muligheter for fremtidig forskning. Integrering av samtaleanalyse med adferdsanalyse kan ytterligere forbedre systemet, noe som kan muliggjøre en mer omfattende forståelse av brukerinteraksjoner.

Preface

This thesis is written as the completion of the 5-year MSc in Cybersecurity and Data Communcation with a specialisation in Digital Economy at Norwegian University of Science and Technology (NTNU). The supervisor of this work has been Professor Patrick Bours at the Department of Information Security and Communication Technology at NTNU. The research presented in this thesis was carried out in collaboration with the company Aiba AS, which specialises in Author Input Behaviour Analysis. A preliminary study for this master thesis was completed in the fall of 2023. The work in this thesis was carried out from January to June 2024.

Acknowledgments

I would like to take this opportunity to thank the special people who have supported me during the work of this thesis.

First and foremost, a special thank you goes to my supervisor, Patrick Bours. His invaluable advice, profound knowledge, and support have been instrumental throughout the course of this thesis.

I would also like to express my deep appreciation to Ståle, Jacob, Fredrik, Bjørn, Tor Håkon, Matija, and Kristoffer for their exceptional camaraderie and support at our beloved office during this semester. Their presence made the work environment both enjoyable and productive. Additionally, I am forever grateful to all my friends in Abakus and Trondheim, who have made my time as a student truly wonderful.

Finally, I would like to extend my heartfelt gratitude to my family for all their support throughout my education.

Contents

Appendices

List of Figures

List of Tables

List of Acronyms

CC Clustering Coefficent.

CIR Conversation Initiation Rate.

eSPD early Sexual Predator Detection.

FDT Fuzzy Decision Tree.

GCN Graph Convolutional Networks.

GMM Gaussian Mixture Model.

NN Neural Network.

SDG Sustainable Development Goal.

TF-IDF Term Frequency-Inverse Document Frequency.

UoI User of Interest.

UoIs Users of Interest.

Chapter 1
Introduction

1.1 Problem Description

The digital landscape is experiencing a significant and troubling increase in online enticement cases. As technology becomes more integrated into our daily lives, the prevalence of predators online exploiting these platforms is growing at an alarming rate. According to the National Center for Missing & Exploited Children (NCMEC), *" Online enticement is a form of exploitation involving an individual who communicates online with someone believed to be a child with the intent to commit a sexual offense or abduction."*[Nat24]. Data from their newest report shows that reports of online enticement surged over 300%, jumping from 44,155 incidents in 2021 to a staggering 186,819 in 2023. Online enticement, which will be referred to as cybergrooming in this thesis, comprises various forms of exploitation. This includes sextortion, where a minor is manipulated or pushed into producing sexually explicit material or engaging in physical encounters with the perpetrator for sexual purposes. The increase in reported cybergrooming incidents is partly due to financial sextortion, in which a perpetrator extorts money from a child by threatening to expose their nude or sexual images to the public.

This rapid increase highlights the urgent need for robust measures to protect vulnerable individuals from online predators, who exploit the anonymity and reach of digital platforms. Understanding the dynamics of these threats and developing effective detection and prevention mechanisms are essential steps in mitigating the risks associated with cybergrooming. Addressing this issue early on is crucial, as it can prevent psychological and physical harm to children. Significant progress has already been achieved in the early detection and intervention of suspicious online interactions before they escalate into dangerous situations. These efforts include a diverse range of strategies and technologies, including automated language monitoring and advanced graph analysis systems.

Previous research has primarily focused on conversation classification. Typically, these systems identify potential predatory conversations by issuing a warning when a conversation surpasses a specific threat level. Other systems have had their primary focus on specific *users*, giving warnings when a user's risk level surpasses a certain threshold. In this project, we will not use a threshold; instead, we will assess the riskiness of users relative to one another. Certain behaviors increase the probability that a user is malicious, resulting in a higher score and a higher rank. Conversely, innocent behaviors lower a user's rank. This approach allows for continuous and dynamic evaluation based on actions, enabling a more accurate distinction between users. As a result, human moderators can prioritize users with the highest ranks, offering a more effective method to manage potential threats on an online platform. The research questions are maintained from the preceding pre-project [Eng23b].

1.2 Research Questions

In this section, "inappropriate" behavior is mentioned alongside predatory behavior, as we expect to detect other unwanted behaviors in our research as well.

Research Question: *Can a user-ranking system based on behavioural analysis streamline the detection of predatory behavior or other inappropriate activities on a platform?*

This research question seeks to investigate methods for detecting predatory behavior online. Unlike previous approaches that score and flag users once they exceed a certain threshold, this study focuses on developing an action-based ranking system that highlights users with the highest scores. The objective is to create a system that is both accurate and effective, aiming to identify potential predators as early as possible with high precision.

Subquestion 1: *What features from the user behaviour will best show the difference between normal and predatory/inappropriate behaviour?*

Determining the optimal features will enhance the system's accuracy and will be beneficial for the subsequent research question to identify the most effective methods for assessing user behavior.

Subquestion 2: *What methods can best be used to evaluate the behaviour of a user such that normal users get scored lower than predatory/inappropriate users?*

This subquestion seeks to identify the best methods, tools, technologies, and techniques to ensure that the ranking system effectively assigns lower scores to normal users and higher scores to those exhibiting predatory behavior.

Subquestion 3: *What would be a good performance metric for this ranking system, where predatory/inappropriate users are ranked higher than normal users?*

Defining a suitable performance metric is crucial as it measures the effectiveness of the ranking system. Selecting a metric that focuses on accuracy while prioritizing the identification of predatory behavior is essential to ensure that the system aligns with the primary objective defined in the main research question.

1.3 Contribution to UN Sustainable Development Goals

This project is highly relevant to UN Sustainable Development Goal (SDG) 16.2, which is the following: *"End abuse, exploitation, trafficking and all forms of violence against and torture of children"* [15]. It also supports the overarching aim of SDG 16 to promote justice and strong institutions. By enhancing our understanding of online grooming patterns, we can equip institutions with the tools needed to safeguard children's rights and ensure their safe participation in digital environments. This research underscores the critical role of technology in advancing global human rights and our commitment to making the internet a safer place for future generations.

Chapter 2 Background

This chapter aims to provide the necessary background information in order to understand the content of this thesis. We have divided it 4 sections, starting of with Section 2.1 covering the act of grooming in general. Section 2.2 provides an in-depth examination of the specific category of grooming known as cybergrooming. Next, Section 2.3 covers the basic principles of graph theory used during this research. Finally, Section 2.4 explains the principles of fuzzy logic.

2.1 Grooming

A 2022 article by Winters et al. [WKJ22] thoroughly evaluated and critiqued previous definitions of grooming. Their proposal for a new operational definition is now well-cited and states the following:
"Sexual grooming is the deceptive process used by sexual abusers to facilitate sexual contact with a minor while simultaneously avoiding detection. Prior to the commission of the sexual abuse, the would-be sexual abuser may select a victim, gain access to and isolate the minor, develop trust with the minor and often their guardians, community, and youth-serving institutions, and desensitize the minor to sexual content and physical contact. Post-abuse, the offender may use maintenance strategies on the victim to facilitate future sexual abuse and/or to prevent disclosure" [WKJ22].

The literature on grooming suggests that it can be seen as a series of stages in which the predator progresses, where the imminent goal is sexual abuse [Lan10] [LPB09]. Resulting from an extensive literature review, Winter and Jeglic [WJ17] suggest 4 different stages of grooming in a 2017 article: victim selection, access gaining, trust development and the sexual stage.

The first stage includes the *selection of the victim.* This could be based on a variety of factors, such as appeal, level of accessibility, and perceived vulnerabilities of the child. Some predators prefer small children over children with a higher perceived level of attractiveness, while others choose their victims based on how they dress [EBK95]. Suppose the victim lives in a single-family household or is in a living situation where parental supervision is limited. In that case, this may increase the chance that the predator chooses this victim. Limited parental supervision could for instance be a result of families struggling with drug or alcohol addiction, mental disorders, or domestic abuse [ODER07] [Fin94]. The predator can also target victims with perceived psychological vulnerabilities, such as low self-esteem and confidence, naivety, or neediness [ODER07]. These psychological vulnerabilities are especially exploited by groomers utilizing the internet to find their victims. A 2023 article by Williams et al. [WEB13] proposes that a child who lacks social support and experiences isolation may be more likely to communicate with strangers online who offer support and acceptance. Taking into account the points mentioned, previous research shows that victim selection is an incredibly strategic process that, in many cases, includes a high level of planning.

After finding a potential victim, the next stage includes *gaining access.* The general goal during this stage is to isolate the victim both physically and emotionally [ODER07]. This can be divided into two main categories; intrafamiliar and extrafamiliar. Intrafamiliar predators already have the approach they need by being a member of the family and can gain access to the victim within the home. Incest predators can, for instance, sneak into the child's bedroom while they sleep. Extrafamiliar predators need another approach to gain access to the child, as they are not a natural member of the child's proximity. Therefore, this type of predator will be in places where children are naturally present, such as amusement parks, malls, and sports arenas [EBK95]. Jobs such as bus drivers, coaches, and teachers could be attractive to a predator of this type, as it places them close to the child both physically and emotionally.

The next stage that Winters and Jeglic [WJ17] propose involves *trust development.* Olsen et al. [ODER07] define this process as the *"the ability to cultivate relationships with potential victims and possibly their families that are intended to benefit the perpetrators own sexual interest"* [ODER07]. At this stage, the predator attempts to establish an exclusive relationship with the child. The common factor is that the predator will try to go from being an acquaintance to becoming an actual friend and confidant of the child [McA06]. For example, by paying more attention, asking about their struggles, or sharing secrets, the child will eventually develop a high level of trust in the predator. Following this, the predator may engage in more peer-like activities, which can depend on the child's age and situation. Such activities can include engaging in conversations about sexual topics or simply playing games if the child is younger. The general goal will be to reach a level of trust where the predators can manipulate the child to participate in sexual abuse at a later stage.

Preceding the sexual abuse, the predator will in many cases gradually introduce the child to physical touch [WJ17]. This stage is called *desensitizing the child to touch*, and can be seen as a preparatory stage for the sexual abuse that the predator will induce on the child later. The first introductions can be hidden as "accidental" or innocent touches, such as hugs, tickling, and pats on the back. Gradually, the predator will initiate activities that involve more physical contact. Playing hide and seek in the dark, nude swimming, strip poker, and wrestling are examples of such activities [McA06] [Lan10]. Berliner and Conte [BC90] suggest that the desensitization stage can be done not only physically, but also psychologically. For example, to achieve increased sexualization, the predator may discuss sexual matters with the child.

2.2 Cybergrooming

If you put aside the aspect of building trust with individuals other than the minor, defined in Section 2.1 [WKJ22], grooming can be termed cybergrooming when utilized to exploit children on digital platforms. The methods of predators that use an online platform to connect with the child are in many ways very similar to the predators that manipulate the child physically. A 2003 article conducted by O'Connell studies the typology of child cybersexploitation and online grooming practices, where she proposes a variety of stages that predators progress through with the victim [OCo03]. Depending on their goals and intents, some people may skip over particular stages entirely or become trapped on them for extended periods of time. O'Connell proposes 5 stages, where every stage is distinguished by goals that tackle psychological elements associated with the predator's assessment of the child's susceptibility. She presents the following stages: *friendship forming, relationsship building , risk assessment, exclusivity, and sexual.*

Due to the nature of connecting with a child online, the stages the predator and the child move through are mostly conversation-based. This implies that the predator must be cautious in choosing which topics to investigate.

The first stage is the *friendship forming* stage, where the predator exchanges high-level information about the child, such as age, gender, and where they're situated. The general goal during this stage is to get to know the child, and the predator usually requests non-sexual pictures of the child to see if the child matches their particular preferences. The predator will also do this to ensure that he's talking to an actual child [OCo03].

The next stage O'Connell proposes is an extension of the friendship forming stage, called the *relationship forming stage.* By showing engagement in the child's interests, home, and school life, the predator will attempt to create an illusion that they are "best friends". This will initiate a deeper, more trusting connection.

After establishing a certain level of trust with the child, the predator will move through a stage called the *risk assessment stage.* During this stage, the predator will assess the likelihood of being detected. Information such as parental supervision and the number of people using the computer/tablet/phone is essential for the assessment.

Closely following the risk assessment, the predator will begin to form a sense of *exclusivity* with the child. Phrases like "I understand what you're going through and you can talk to me about anything" are common in this stage as is creates a strong sense of mutuality. The predator will portray him or herself as a confidant who understands the child uniquely, encouraging the child to keep their relationship secret.

Eventually, the conversation may reach the *sexual stage*, which is the final stage when in the context of "cyber"-grooming. The predator can introduce this step in many different ways, but questions in the area of "have ever you tried touching yourself" or "have you ever kissed someone" are typical. During the previous stages, the predator and the child have developed a deep connection, making the intention of these types of questions appear more innocent. The sexual stage can also introduce the exchange of sexual pictures.

A study conducted in 2022 by Rezaee et al. [RRB22] examined all research papers on grooming detection by studying the psychological definitions and facets of grooming. Figure 2.1 shows their proposed taxonomy for online grooming detection problems [RRB22].

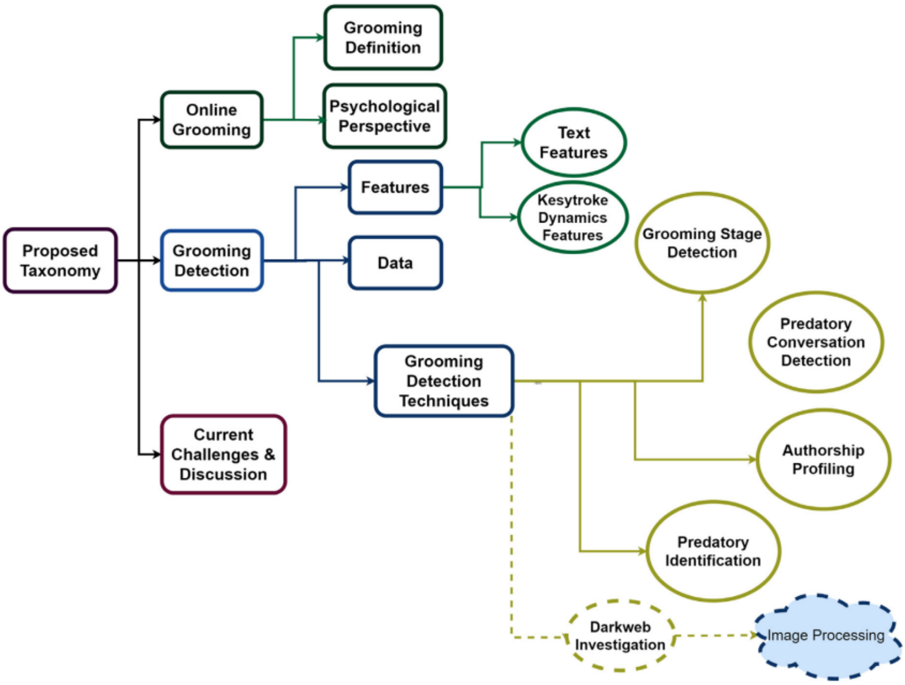


Figure 2.1: Proposed taxonomy for online grooming detection problems [RRB22, Figure from page 3]

A 2011 study by Briggs et al. [BSS11] explored internet-initiated sexual offenses and chat room sex offenders. They discovered that online predators can be divided into two subgroups; fantasy-driven and content-driven. The fantasy-driven predator prefers to engage in sexual activities online, such as sexting, without actually intending to meet the child offline. Sexting refers to the act of exchanging sexually suggestive photos or messages using mobile devices such as cell phones and other forms of mobile media [GBGZ13]. On the other hand, content-driven offenders are motivated to engage in physical sexual behavior with the child, where the online activities are seen as a means to facilitate this. Table 2.1 displays some of the features of contact-driven and fantasy-driven predators presented in Briggs et al. study [BSS11]. The study was based on a sample of 51 convicted chat room sex offenders.

Feature	**Content-driven predator**	**Fantasy-driven predator**
Demographics	Young, often single, low education, higher unemployment rate.	Older, many married or divorced, low unemployment rate.
Online Sexual Behaviors	Grooming for physical meeting, few other sexual activities online.	Diverse cybersex behaviors, explicit sexual communication.
Mental Health	Less frequently diagnosed with a paraphilia.	More often diagnosed with a paraphilia and narcissistic personality disorder.
Compulsive behaviour	Compulsive pornography use.	Compulsive pornography use.

Table 2.1: Features of content-driven and fantasy-driven predators [BSS11]

2.3 Graph theory

Graphs and networks are utilized across various fields to depict connections between entities, including social interactions among people, connections among web pages, and traffic flows [TZHH11]. This section covers graph and network theory used during our reserach.

2.3.1 Graph basics

Agarwal et al. [AS09] defines a simple graph G as a pair $G = (V, E)$, where

- V represents a finite set known as the nodes or vertices of G, while
- E represents a set of unordered pairs of nodes. The elements of E are referred to as the edges of G.

E can also be represented as the relationship between the nodes $E(\subseteq V \times V)$.

Graphs can possess specific details, such as directionality, which is introduced through undirected and directed graphs. An undirected graph G is made up of a set V of nodes and a set E of edges, where each edge e in E connects a pair of nodes without any specific order. An edge joining the node pair i and j can be referred to as either i, j or j, i. In simpler terms, an undirected graph can be described as a graph without any start and end point on each edge.

On the other hand, graphs can be directed. A *directed graph* G is composed of a set V of nodes and a set E of edges, where each edge $e \subseteq E$ is connected to an ordered pair of nodes. Each edge has a specific direction in a directed graph and in a diagram, each $e = (u, v)$ is depicted by an arrow. This can be observed in figure 2.2.

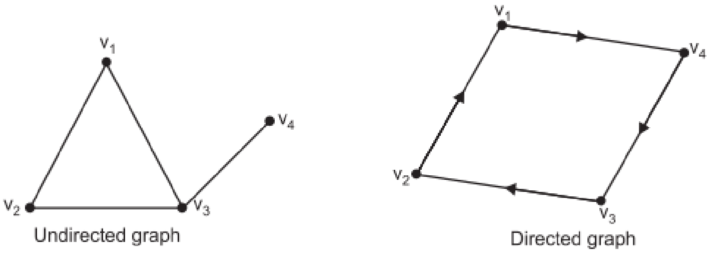


Figure 2.2: Directional graphs [AS09]

In certain graph applications, knowing the relationship between different nodes is central for conducting graph analyses [TZHH11]. For instance; how many messages have two people exchanged in a chatroom or how much traffic has passed through a node in a network. These relationships can be represented through weights, where an edge is assigned a numerical value. This type of graph is called a *weighted graph*. It is denoted as triple $G = (V, E, w)$, where V is a set of nodes, E is a set of edges, and $w :\rightarrow E \rightarrow \mathbb{R}^+$ assigns a weight to each edge $e \subseteq E$ [TZHH11]. Another feature that could be added to the edges is a label, which is when edges are labeled with name or data. This is called a *labeled graph* [AS09].

If an edge e connects two nodes v and u in a graph G, they are said to be *adjacent*, or in simpler terms, neighbors. The edge is then said to be *incident* to those nodes.

A convenient method to represent graphs is by using an adjacency matrix $A \in \mathbb{R}^{|V| \times |V|}$. To represent a graph with an adjacent matrix, it's necessary to organize the nodes in the graph so that every node corresponds to a specific row and column in the matrix. The presence of edges is indicated as elements in the matrix: $A[u, v] = 1$ if $(u, v) \in E$, and $A[u, v]$ otherwise. A graph containing only undirected edges will result in a symmetric matrix. When representing weighted graphs, the values in the matrix are values decided by the weight of each edge, hence not restricted to be 0,1 [Ham20].

In a directed graph G, the out-degree of a node V, is denoted as $deg^{+}(V)$ and represents the number of edges beginning at node V [Ham20]. The in-degree, denoted $deg^{-}(u)$, represents the number of edges that end at node V. By taking the sum of $deg^{+}(u)$ and $deg^{-}(u)$ you will get the *total degree* of the node, also called the *node degree.* With the adjacent matrix A, equation 2.1 is used to calculate the degree of the node.

$$\deg(u) = \sum_{u \in V} A[u, v] + \sum_{v \in V} A[v, u] \tag{2.1}$$

Previously mentioned was the concept of weighted graphs. In some cases, it is highly relevant to know both the weighted in-degree and out-degree. This, for instance, could represent the total amount of messages a user has received from a variety of different users. The calculation for weighted out degree can be performed by using equation 2.2.

$$\deg_{w}(u) = \sum_{v \in V} w_{u,v} \tag{2.2}$$

2.3.2 Sub graphs and ego graphs

Another important part of graph theory that is very relevant in this thesis are *sub graphs.* A graph g is considered to be a sub graph of another graph G if all the nodes and edges in g are also present in G [AS09]. A sub graph can be considered as being within or a component of another graph, as implied by its name. *Ego graphs* are a specific type of sub graph, that will be highly utilized when we analyze networks. This type of sub graph consists of a single node u and the network between u and the edges between u and v_i [LM12]. We refer to node u as the *ego* and v_i as *alters.* Put differently, the ego graph allows us to identify both nested and overlapping clusters within the ego network of u. A simple ego graph is shown in Figure 2.3.

.

[1] A simple ego-graph [Image]. Retrieved June 1, 2024.

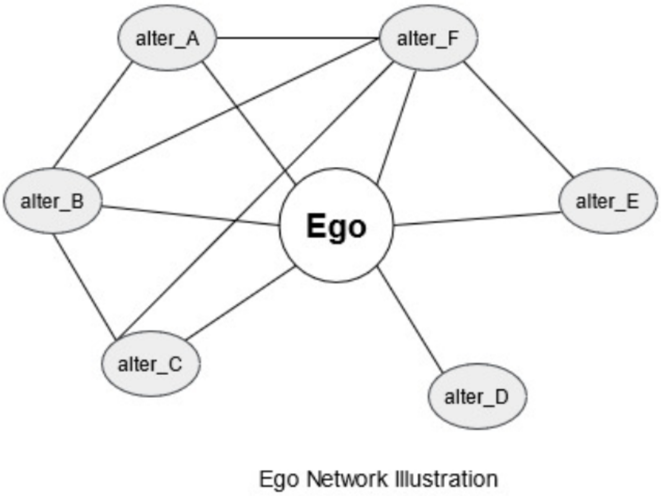


Figure 2.3: A simple ego-graph[1].

Clustering Coefficient (CC)

After obtaining a sub graph or an ego graph for a graph G, it can be useful to have a measurement of how interconnected the network formed is. The CC is a useful measure because it quantifies how often the neighbors of a node are also connected to each other, effectively measuring the degree to which nodes in a graph tend to cluster together. Imagine your group of friends. The CC measures how many of your friends are friends with eachother. A CC value of 1 means that every friend of yours are also friends of eachother. This can be extremely valuable when analyzing the social networks of users online. Equation 2.3 shows how the local variant of the CC is computed [Ham20] [Eng23a]:

$$CC(v) = \frac{2N(v)}{\deg(v) \times (\deg(v) - 1)} \tag{2.3}$$

The numerator in equation 2.3 counts the number of edges between neighbors of node u, while the denominator calculates how many pairs of nodes there are in u's neighborhood. The factor of 2 in the numerator accounts for the fact that each edge between neighbors is counted twice, once for each node it connects. Figure 2.4 shows three different graphs with different CC values.

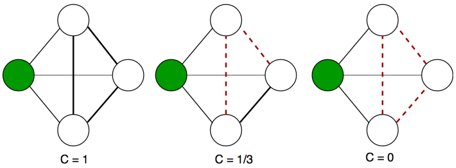


Figure 2.4: The figure shows different ego graphs with various CCs [2].

2.4 Fuzzy Logic

Fuzzy logic, first introduced by Lotfi A. Zadeh in 1965 [Zad65], represents a significant advancement in logical systems by addressing the nuances of approximate reasoning. Unlike classical two-valued logic systems, which classify propositions strictly as true or false, fuzzy logic allows for degrees of truth. This means that propositions can have a truth value that is not only binary but can take any value within a given range, such as the unit interval [0, 1]. This flexibility is important for representing and reasoning about the types of nuanced and imprecise information frequently encountered in real-world situations.

One of the primary features that distinguishes fuzzy logic from classical logical systems is its ability to handle fuzzy predicates and fuzzy quantifiers. In classical logic, predicates must be precise, meaning they can only be completely true or completely false, with no in-between. In contrast, fuzzy logic allows predicates to be fuzzy, such as "tall," "young," or "hot," which do not have sharp boundaries. Additionally, fuzzy logic introduces fuzzy quantifiers, such as "most," "many," and "few," which can be used to describe imprecise quantities and frequencies. These fuzzy quantifiers provide a way to represent and manipulate probabilities within logic, thereby enhancing its expressive capabilities [Zad65].

Furthermore, fuzzy logic includes various modes of qualification for propositions, such as truth-qualification, probability-qualification, and possibility-qualification. This enables a finer and more adaptable method for evaluating propositions. For example, a proposition can be qualified as "not quite true," "unlikely," or "almost impossible," reflecting the varying degrees of certainty we

[2] Clustering Coefficient [Image]. Retrieved June 1, 2024.

often experience in practical scenarios. By addressing the limitations of classical logic systems, fuzzy logic provides a robust framework for modeling and reasoning about the complexity and ambiguity inherent in many real-world problems [Zad65].

The best way to illustrate this is through a simple example, provided in Figure 2.5.

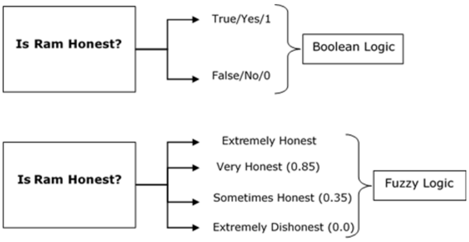


Figure 2.5: Introduction to Fuzzy Logic.[3] Tutorials Point. (n.d.)

In this paper, we do not delve deeply into the mathematical foundations of fuzzy logic, as our primary focus is on utilizing its concept of degrees of truth to address the nuances of predatory behavior. The detailed mathematical aspects are not central to our discussion and are therefore omitted. We will utilize the term FDT when discussing relevant aspects of fuzzy logic. An example of a full FDT can be seen in Figure 2.6, where the main goal is to decide whether to play volleyball or not based on a combination of weather conditions.

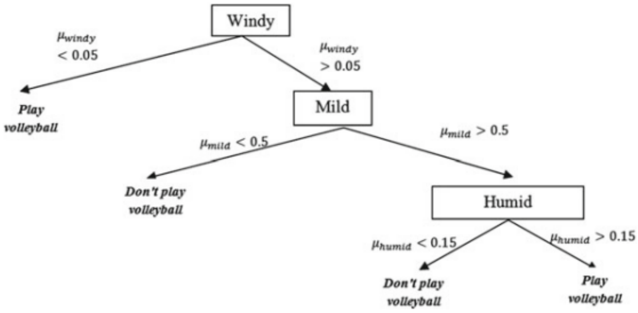


Figure 2.6: FDT on playing volleyball [AC16, Figure on p. 256].

[3] Introduction to Fuzzy Logic [Image]. Retrieved June 1, 2024.

Chapter 3 State of the art

This chapter serves as an introduction to the latest relevant studies on early detection methods, as well as studies that utilize graph theory and anomaly detection to detect cybergroomers. It is built upon the most important and relevant papers found and researched in the pre-project preceding this thesis [Eng23b].

3.1 Early detection

In this section, we will review a selection of state of the art papers focused on early detection methods.

3.1.1 Detection of Cyber Grooming in Online Conversation

In 2019, Kulsrud and Bours [BK19] conducted a study to detect sexual predators in online chat conversations using three different analytical approaches: message-based, author-based, and conversation-based. Each approach was combined with five classification algorithms and two feature sets to identify the most effective method.

Their study revealed that the best results were achieved with the author-based approach when using Neural Network (NN) classifiers, and with the conversation-based approach using Ridge or Naïve Bayes classifiers. Both of these approaches utilized the Term Frequency-Inverse Document Frequency (TF-IDF) feature set, which proved to be the most optimal feature for predator detection. TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents, making it highly effective for text classification tasks.

The author-based approach with NN classifiers effectively identified predators by analyzing all messages from a single user within a conversation, while the conversation-based approach excelled by considering the entire conversation context to classify the chat as normal or suspicious, and then identifying the predator among the participants in suspicious conversations.

3.1.2 Early Detection of Sexual Predators in Chats

In 2021, Vogt et al. [VLA21] took on early Sexual Predator Detection (eSPD) through chat analysis. They emphasized the importance of raising alerts as early and accurately as possible, creating a comprehensive evaluation setup using state of the art BERT-based language models at that time. Their system continuously analyzed each new message in ongoing chats, dynamically updating the user's risk score based on the content and context of the conversation. When a user's cumulative risk score surpassed a predefined threshold, an alert was triggered, signaling potential predatory behavior. Their study defined a "warning latency" as the number of messages exchanged before the system could trigger a warning.

To test the effectiveness of their eSPD system, Vogt et al. tested three different BERT models: BERTlarge, BERTbase, and MobileBERT. Among the tested models, BERTbase provided the best overall performance, balancing precision and recall effectively.

3.2 Graph theoretical approaches

In this section, a selection of the state of the art papers using graph theory will be discussed.

3.2.1 A graph theoretical approach to online predator detection

In 2022, Aarekol [Aar22] conducted a study using a graph theoretical approach to online predator detection. Given a large dataset from an online game, she utilized a number of tools and methods to analyze how different users behave in the network. The Pyvis Python library was used to visualize the graph representation of the network. Subgraphs in the form of ego-graphs, which was explained earlier in chapter 2, were then extracted from the networks, allowing for better interpretation of different user behaviors. These graphs and the information within reveal a variety of features that are highly important and relevant when analyzing graphs. The Python library NetworkX was utilzed to calculate the values of these features, where a selection of the most important features are CC and the distribution of the length of the conversations between a user and its neighbors. Some examples of the node-specific features are:

- The number of messages sent to or from node n (weighted degree).
- The number of neighbors who have sent messages to node n (in-degree)
- The number of neighbors who have received messages from node n (out-degree)

In total, 22 feature values were calculated for each node. The set of features was calculated for all nodes and formed a feature vector used in the clustering algorithms. The algorithms tested for predator detection were agglomerative clustering, BIRCH, k-means, mean shift, DBSCAN, and Gaussian Mixture Model (GMM), which were programmed using the Scikit-Learn library in Python. The algorithms are executed by programs that read a CSV file containing nodes represented as a feature vector. The clustering scripts normalize the features and perform clustering using Scikit-Learn. Several of the algorithms detected a small number of predators and disclosed illegal activities in the chats, where Birch was a particularly good algorithm.

An important observation from the results was the occurrence of high percentages of low-weighted conversations in the predator users, implying that the predator contacts many users without having longer conversations. Highly related, the out-degree divided by the total degree, could detect users who frequently engage in conversations where the other person does not respond. This is also common among heavy spammers. Also notable from the results was the fact that some predators were mostly active in shorter periods of time, which is harder to detect in datasets that have large time frames.

Our system will also take a graph-theoretical approach, particularly taking advantage of ego-graphs and the features they provide. However, in contrast to Aarekol's work, which analyzes the datasets that have taken place over an extensive period of time, our system will be *action-based.* This means that an updated evaluation of a user will take place after every new action of that user. The user's behavior will be analyzed in a live ranking system, where their riskiness will only be compared with other users. Aarekol employs traditional clustering algorithms in her study, that can help identify patterns or groups of users exhibiting similar behaviors. However, these algorithms can require a lot of processing power, especially in scenarios requiring real-time processing and for large datasets. They typically require access to the entire dataset to form these clusters, which might not be possible in a live system like ours, where data is constantly changing.

3.2.2 Dynamic graph theoretical analysis of cybergrooming detection in chatrooms

A 2023 thesis by Eng [Eng23a] uses a somewhat similar dynamic graph-theoretical analysis to detect cybergrooming in chatrooms. Her first steps after cleaning and filtering the data, were to identify behavior that would distinguish normal users from abnormal users. This, among other things, involved analyzing node behavior with ego-graphs, like [Aar22] did in her thesis. By using the NetworkX python library, the goal was to visualize how ego graphs developed over time. She discovered that there is a clear difference when it comes to the interconnectivity between neighbors of the normal user and the abnormal users. The normal user's neighbors have a significantly greater amount of connections in-between them compared to the abnormal user's neighbors. Aarekol's [Aar22] thesis made references to this as well. The ego-graphs can be seen in Figure 3.1 and Figure 3.2, showing how their social networks of two users develop over time.

She also noted, much like Aarekol [Aar22], that it became apparent that the frequency of messages was important to extract from the simulation. Over a few minutes and hours, the normal user made contact and exchanged messages with a handful of different users, who also interacted with each other. However, abnormal users had a slower message exchange pattern, which could be a consequence of the other person being cautious when talking to a stranger. As mentioned earlier, most neighbors are not conversing with each other in this user's ego graph. Based on the data extracted from the graph simulations, she chose a variety of different features to monitor the users' behavior over time. In addition, context-specific (conversation-specific) features of interest were added, such as response time and time since the last activity.

Such features were extracted using the DiGraph class in the NetworkX library, which created a directed one-hop neighborhood resembling the users' interactions.

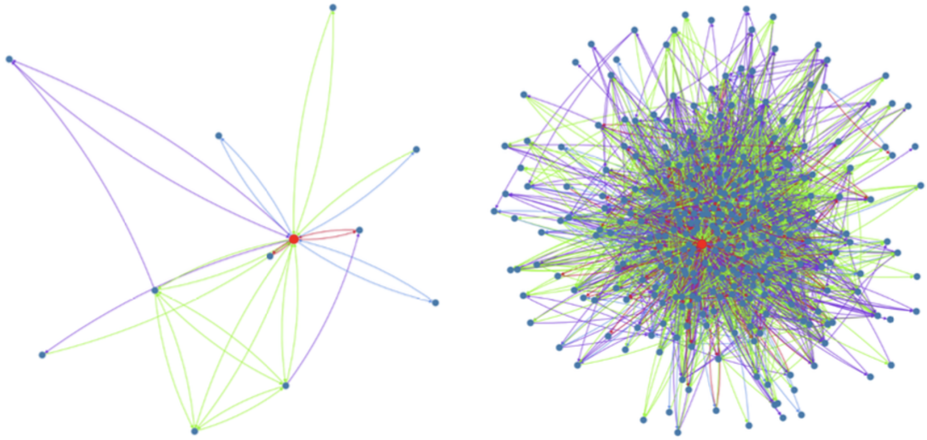


Figure 3.1: Two ego graphs of the same normal user, captured after a few interactions and after a substantial number of interactions [Eng23a, Figure on page 36].

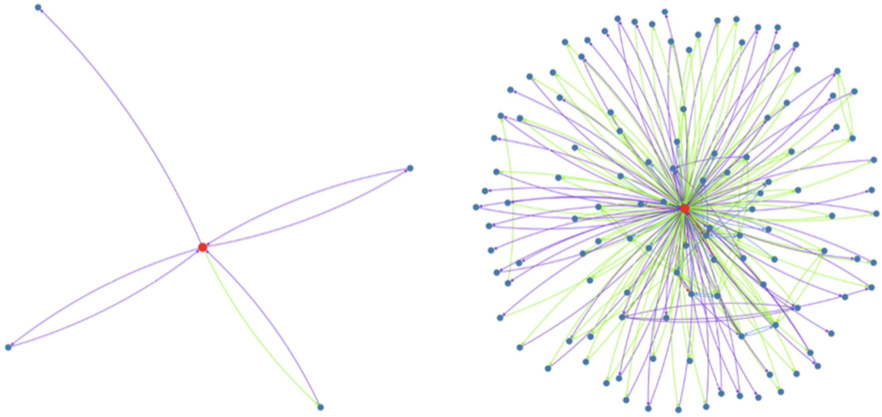


Figure 3.2: Two ego graphs of the same abnormal user, captured after a few interactions and after a substantial number of interactions [Eng23a, FIgure on page 37].

To monitor user behavior and identify abnormal patterns, they employed the ML-classification algorithm SVM with a linear kernel using the SciKit Python library [BB19]. For testing, different features were grouped to test various combinations. A 5-fold cross-validation with an SVM model was then used to assess the performance of the different feature combinations. The evaluation metrics utilized were accuracy (ACC), precision (PR), recall (RC), and the F_β-score, where $\beta = 0.5, 1, 2$ indicates the weight of recall in the combined score.

The detection mechanism operates by collecting the probability scores for user behavior as messages are sent. To ensure data sufficiency, it initiates the analysis after at least 200 messages have been sent. Next, it examines the user's probability scores to identify a period where they stabilize or exhibit consistency, demonstrating a stable behavioral pattern. Comparing these scores to a predetermined *threshold*, the detection mechanism classifies the user as either normal or abnormal based on the behavioral pattern.

The results from Eng's thesis revealed several important behavioral patterns to consider in our study. An important observation was the difference in the frequency of messages and the number of messages between normal and abnormal users. Abnormal users often exhibited a higher amount of conversations that consisted of less than five messages exchanged, and the interconnectivity between their neighbors was also minimal. Eng suggests that future research into cybergrooming detection should focus on features such as the CC and message distribution when using graph-theoretical approaches [Eng23a].

The simulation of the ego-graph also revealed that changes in activity levels could serve as a meaningful indicator. Abnormal users, for instance those operated by spammers, were marked by brief, yet intense periods of communication. Such patterns of behavior could be characteristic of certain cybergroomers, as these may engage in cybergrooming at any given opportunity to fulfill their needs. To accelerate the process, they may log in, and simultaneously try to connect with multiple users [Eng23a].

Our study will use several of the features introduced in Engs thesis, but will also have significant differences. As mentioned previously, our system will be live and action-based, updating the ranking of users continuously. Thus, the use of SVMs will not be possible because our system will not include a threshold mechanism. However, many of the node-specific features introduced in Engs thesis will be included when performing the risk assessment, as well as tools like NetworkX [Eng23a].

3.3 Anomoly detection

Anomaly detection aims to identify instances in datasets that differ significantly from the norm [CC19]. A 2021 paper from Kumagai et al. [KIF21] investigates anomaly detection in attributed graphs, which are graphs where the nodes and edges have additional attributes beyond the structural connections. In anomaly detection in social networks, these attributes can be seen as the connection of social relationships. The authors propose a semi-supervised anomaly detection framework that utilizes Graph Convolutional Networks (GCN) to embed nodes, by leveraging both their attributes and the graph's structural information. Here, semi-supervised means that some instances are labeled for training purposes. The GCN were trained to find abnormalities within the network in real-world attributed datasets, generating unique embeddings (vectors) for each node. By measuring the distances from these node embeddings to the center of a hypersphere, which represents the normative data pattern, it effectively identified outliers. This technique outperformed several established anomaly detection methods in performance.

In our system, we can take inspiration of the use of a semi-supervised learning approach, which leverages a small number of labeled instances to guide the learning process. This could be analogously interpreted as employing a labeled set of predators or predatory behavior, along with unlabeled data as well.

Chapter 4 Data Preprocessing and Analysis

This chapter provides a extensive overview of the preprocessing steps and initial analysis conducted on a dataset. The dataset, consisting of messages exchanged over a three-month period, required significant preprocessing to ensure usability and privacy protection. We detail the methods used to read and concatenate the data, followed by the labeling process where we marked predatory behavior. Additionally, we present a preliminary data analysis and statistical examination of user interactions within the network.

4.1 Preprocessing

The data set that we use in our study is collected from an online game, where the target player base is children. The game, which is anonymized, revolves around interacting with other players while playing the objectives of the game. Private chats and group chats are an important part of the game, where players can get to know each other and negotiate trades for different items. Many players use these chats to trade personal information, even though this is against game policies. Therefore, game developers have added filters in the chat that restrict the exchange of personal information, as well as limiting the use of sexual and harrasive wording. However, players still seem to find ways to work around these filters by using slang language and creative language with punctuation and special characters.

The data provider for this study consisted of a data set with a 3-month timeframe from 01/08/2022 to 31/10/2022. It came in the format of 81 separate `.parquet` files, which needed further processing to be readable. A script was created to read multiple Parquet files from a specified folder into dataframes using the Pandas[1] python library, then concatenating these dataframes into a single datagrame. The dataframe consisted of the following columns: `dateUtc`, `messageId`, `context`, `gameId`, `initiator`, `receiver`, `content`. The first 10 rows of the dataframe can be seen in Table 4.1.

[1] Pandas documentation

Date	MessageID	Context	GameID	Initiator	Receiver	Content
2022-08-22 21:31:20.203	1119	53B5	5**i	C7FF	087A	I LOVE U
2022-08-22 21:31:20.237	2FD9	9938	5**i	F193	E1C7	Addison
2022-08-22 21:31:20.305	FCF6	9CCC	5**i	F674	4FBD	bad*
2022-08-22 21:31:20.292	312E	C31C	j**d	CBEA	FB36	Wait, u mean irl or in the game?
2022-08-22 21:31:20.356	B4F5	4D8A	5**i	A26F	5BD2	hello! is there anything specific you're looking for?
2022-08-22 21:31:21.049	4812	7EBE	y**u	A2B4	8D61	k.
2022-08-22 21:31:21.456	160C	7463	5**i	986D	6BDC	Hey! would you be interested in joining a disc...
2022-08-22 21:31:21.784	FBAF	397A	j**d	BF07	5734	boo
2022-08-22 21:31:21.820	12E7	E59B	5**i	C497	DEFA	(square_clown)
2022-08-22 21:31:21.918	F50E	1B08	5**i	A76E	D1B4	cookie hair?

Table 4.1: The first 10 rows of the datasets, with shortened anonymized IDs for privacy reasons.

All identifying information, including the initiator and receiver IDs, has been anonymized within the dataset. This anonymization ensures the preservation of user privacy, as potential predators or other sensitive entities can only be discerned through their randomized user IDs.

4.2 Data Selection and Labeling

An important part of the preprocessing and preparation of the data set is labeling. This includes annotation of known instances of predatory behavior. Due to the size of the dataset, we extracted 1000 random contexts into a CSV file, with the criteria that they consisted of at least 5 messages. A short extraction of the CSV file is shown in the snippet below, where the header consists of the `'context'`, `'label'` and `'count'`.

```
context,label,count
08A2,0,18
AD3B,0,17
A37A,0,59
BD83,0,11
```

This will be referred to as the *primary dataset* in this chapter. The context is the unique contextID for each conversation, while the count is the total number of messages in the conversation. The label was set to 0 by default and changed to 1 if we found the conversation to be predatory or sexual. To label a conversation as interesting (1) in our case, the conversation needed to have some kind of sexual nature between the two parties. This could, for instance, be initiation of sexting or engaging in roleplaying with intent of a sexual nature. In addition, any indications of abnormal behavior that indicate that an adult user seeks out younger users were labeled 1. To avoid bias, a collaborative effort was established with another group working on the same dataset. Whenever uncertainty arose, double-checking and discussion of labeling decisions were conducted to ensure accuracy. An example of a conversation we labeled 1 is shown in Table 4.2.

For our study specifically, we also annotated the *initiator* of topics of sexual nature, which was displayed when we individually examined the content of each of the unique context. These will be referred to as Users of Interest (UoIs) to aid in further analysis. In Table 4.2, the User of Interest (UoI) would be the user

Sender	Content
EB0D	uu vute??
F7F8	ty
EB0D	how old ar u?
F7F8	15
EB0D	you're a little small and sweet??
F7F8	huh

Table 4.2: Snippet of a conversation label as 1, with shortened userIDs in the first column. It has indications of an adult initiating a conversation with a child.

initiating the conversation. As a final step before the actual data analysis, we created a new file containing all the users involved in general contexts, which is the unique receivers and initiators associated with those contexts. Similarly, all the user we denoted as UoIs were extracted into a separate file containing only the userIDs.

4.3 Datastudy

As mentioned previously, we had access to a dataset encompassing a three-month period of conversations. The dataset comprised 214,848 distinct users who engaged in 1,323,830 distinct conversations. The average conversation length was calculated at 19.72 messages, with a median value of 4.0. Notably, the longest conversation in terms of message count contained 101,668 messages, while the lowest contained 1 message.

4.3.1 Ego graphs

Ego graphs can be essential to understand the behavior of the users. We chose one normal user and one UoI based on the labeling, and constructed their ego graphs using the *NetworkX* library[2]. Figure 4.1 shows the ego graph of *'userX'* who exhibited non-predatory behavior.

The edges within the network are additionally colorcoded based on their respective weights. Specifically, if the interaction between the ego and its neighbor comprises 5 or fewer messages, the edge is represented in green. For interactions

[2] NetworkX documentation for Ego Graphs

involving 6 to 15 messages, the edge color is displayed in blue. Similarly, interactions ranging from 16 to 30 messages are denoted in orange, while interactions exceeding 30 messages are depicted in red.

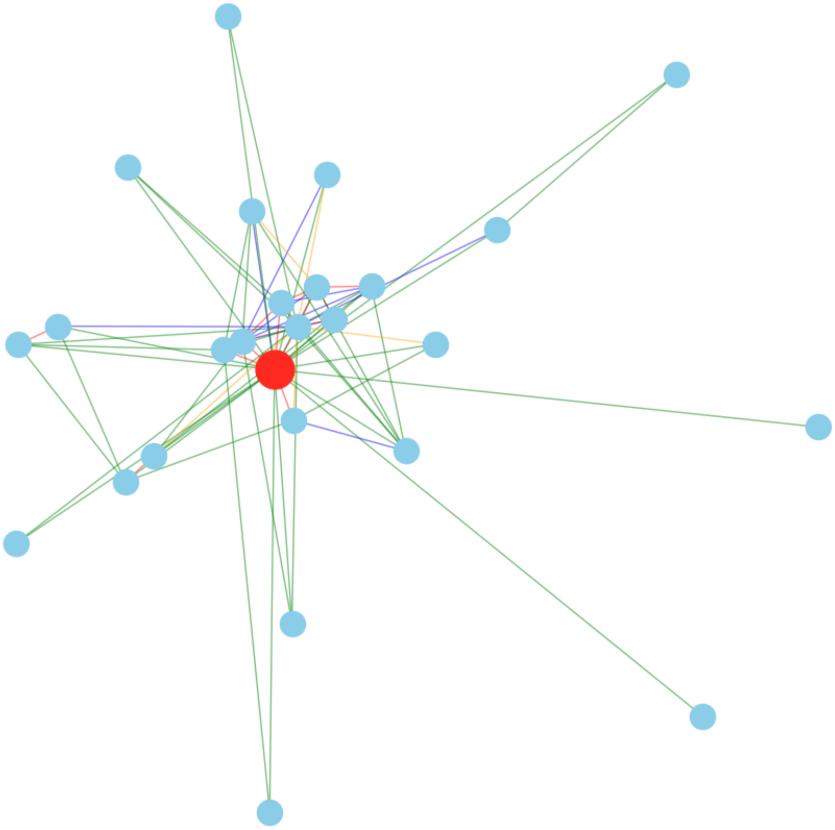


Figure 4.1: Ego graph for (a normal) 'userX'.

You can observe that the user's neighbors are highly interconnected by looking at the edges connecting them. In fact, this particular user is found to have only two neighbors that are not interconnected with the remainder of their social network. Another important observation is that the user has multiple red and orange edges, indicating several conversations of substantial length. This is typical for a user that's part of a bigger friend group, such as a school class or sports team.

On the other hand, we have *'userY'* who depicts more predatory-like behavior. This UoIs ego graph is shown in Figure 4.2.

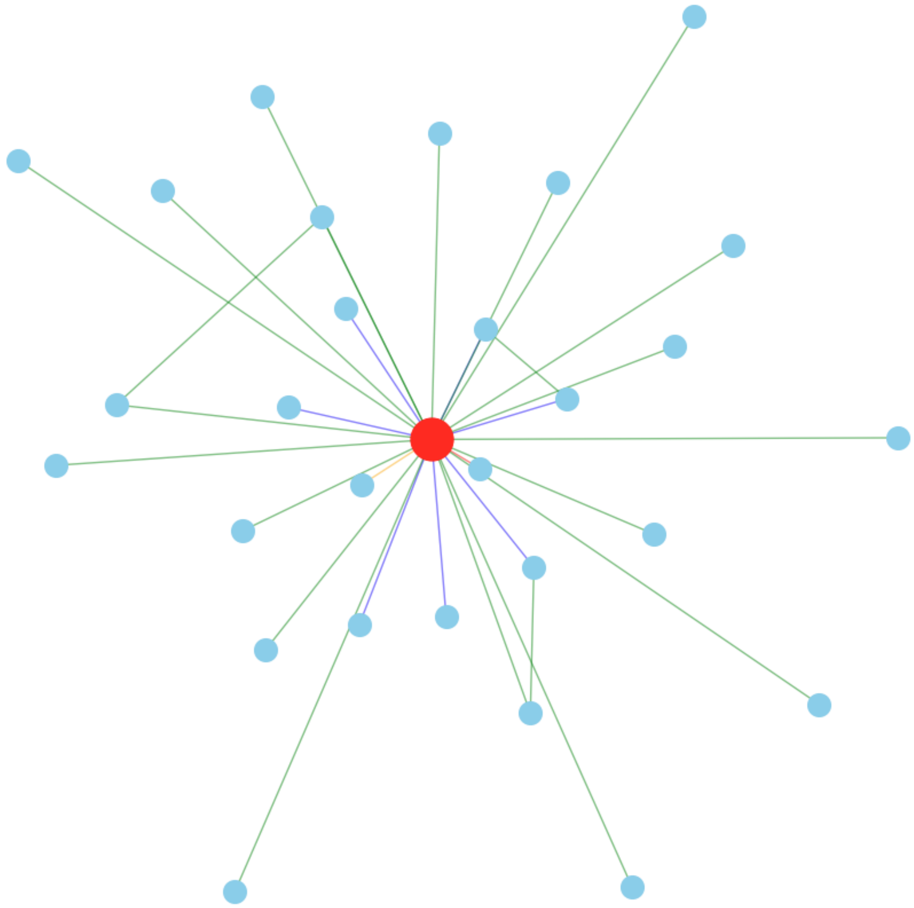


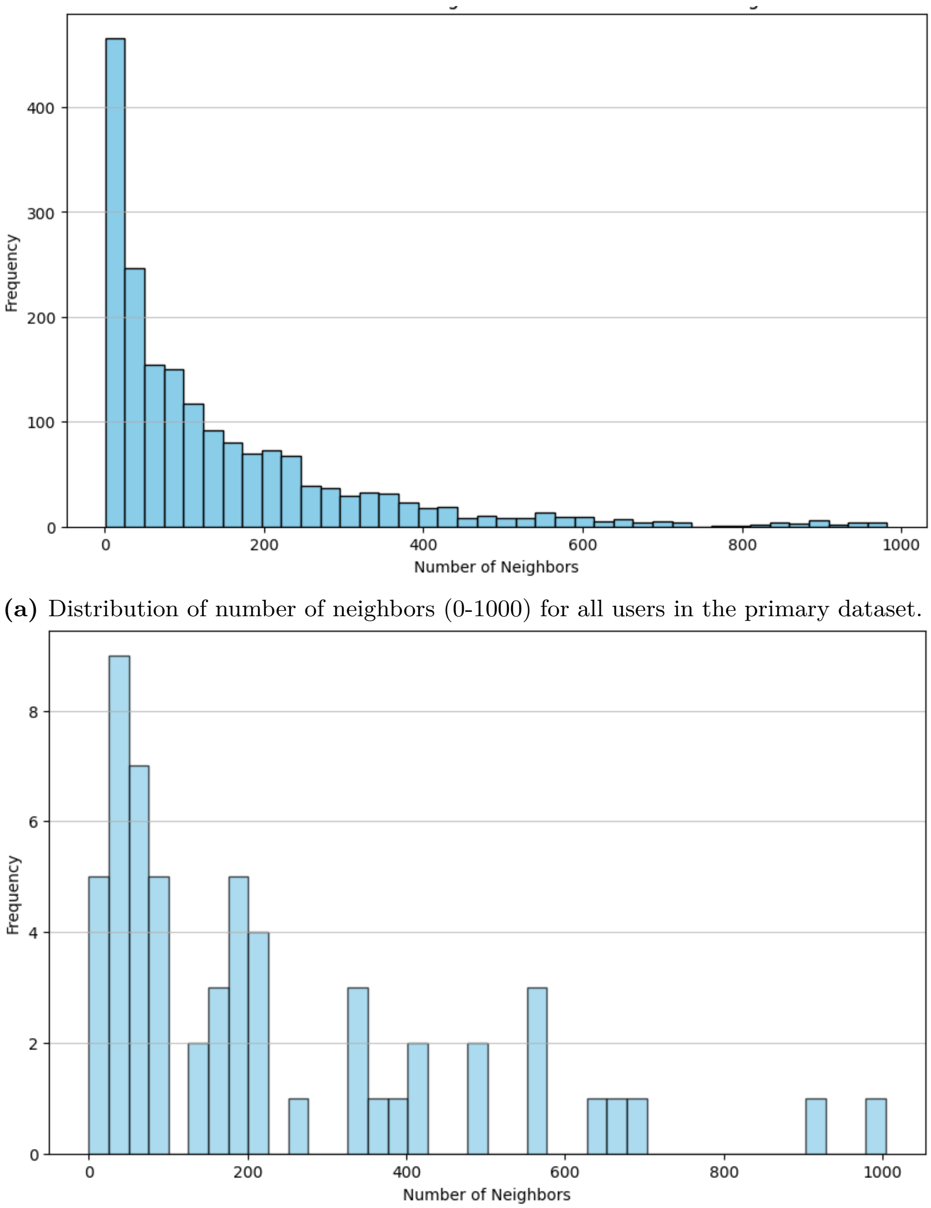
Figure 4.2: Ego graph for (UoI) 'userY'.

This user's social network is significantly less connected than the normal user's, in fact, less than 5 connections between neighbours can be seen. The CC for this user would be significantly lower in comparison to the ego graph of the normal user. Furthermore, there is a noticeable increase in the proportion of green edges, representing conversations consisting of five or fewer messages. This trend may signify the user's difficulty in establishing substantive dialogues and relationships within the community, which is often typical for predators.

Based on the users in the primary dataset, we also got important insights regarding the number of neighbors for the nodes. We discovered this by utilizing the node degree method associated with the Graph class[3] in NetworkX, which will be further explained in the next chapter. The average amount of neighbors across nodes in the primary dataset was 152.2, with a corresponding median of 86.0. On the other hand, UoIs averaged 226.5 neighbors, accompanied by a median value of 156.0, signifying a notable deviation from the overall trend. The distribution of neighbors in the primary dataset of any label can be seen in Figure 4.3a, while the UoIs' distribution is displayed in Figure 4.3b. We have limited the graphs to include a maximum number of 1000 neighbors, since only 9 outliers exist.

Figure 4.4 provides a comparison of the distribution of neighbors based on percentage. The few extremes with over 1000 neighbors have been excluded from the data here, as they are not deemed significant for the interpretation of this chart. A noticeable trend is evident in the distribution, indicating that UoIs tend to exhibit a lower percentage of neighbors in the lower range compared to normal users, while displaying somewhat higher percentages in the higher range of neighbor counts.

[3] NetworkX documentation for the Graph class



(a) Distribution of number of neighbors (0-1000) for all users in the primary dataset.

(b) Distribution of number of neighbors for **UoIs**.

Figure 4.3: (a) Distribution of number of neighbors (0-1000) for all users in the primary dataset. (b) Distribution of number of neighbors for **UoIs**.

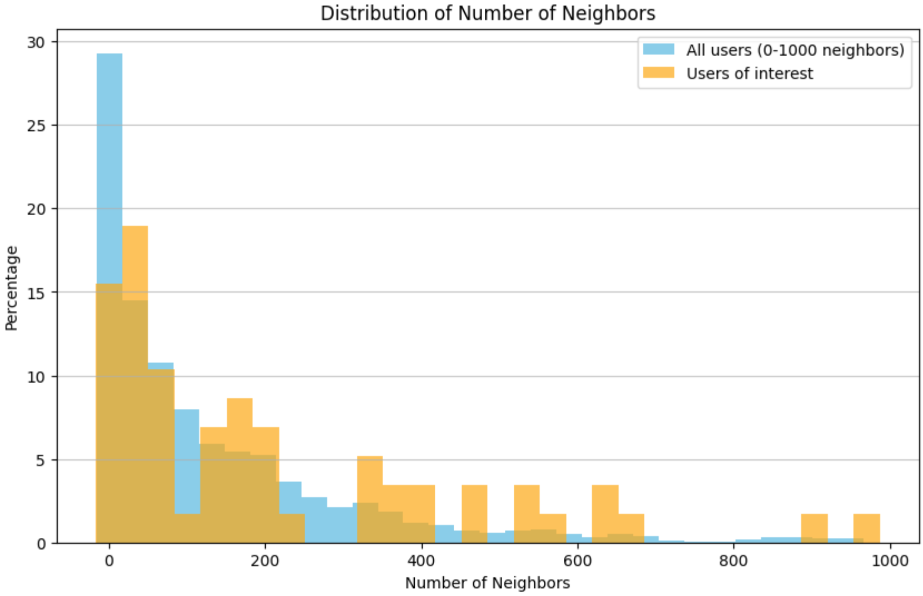


Figure 4.4: Comparison of distribution based on percentage. This chart applies only for users with 0-1000 neighbors, as the "extremes" have been excluded.

Chapter 5
Feature Extraction

This section covers how we methodically identified features indicating predatory-like behavior. The features tested were partially inspired by the work of Aarekol [Aar22] and Eng [Eng23a], who explored this topic in their respective studies.

5.1 Graph utilization

From the work in section 4.3, we learned that it became more efficient and easier to conduct an analysis of the dataset when conceptualized and structured as a network model. As mentioned before, we created separate CSV files for all the users associated with contextIDs in the 1000 contexts dataset. One file contained all the userIDs linked to the contexts in the primary dataset, while the other file specifically contained the userIDs of the UoIs identified as potentially predatory during the labeling process. They consisted of, respectively, 1826 and 58 users. We considered this as essential, as this allows us to use the userIDs to calculate average feature values for the two sets of users.

Initially, we created an undirected graph for the whole dataset using the NetworkX class *Graph*. For each row in the dataframe (shown in 4.1), which consists of an unique `'initiator'` and `'receiver'` connection, an edge was added to the graph G. This type of graph is sufficiently detailed to extract uncomplicated information, such as the average node degree and the CC of the different groups of users. However, we quickly discovered that a *MultiDiGraph*[1] was a better option, as it is a directed graph class that can store multiple edges between two nodes. Each edge can additionally hold optional data or attributes,

[1] NetworkX documentation for MultiDiGraph

which we found to be the edge weights in our case. This was constructed partly similarly as the graph earlier, iterating over each row in the dataset. However, it involves two distinct iterations. First, it computes the number of messages from the initiator to the receiver. Second, it calculates the messages from the receiver back to the initiator. These iterations contribute to the assignment of the ‘weight’ attribute to the two edges.

Figure 5.1 shows a sub graph of the NetworkX Graph consisting of the first 100 nodes in the dataset. Due to certain limitations of the plotting capabilities of the NetworkX library, it does not show the multi-edges of the graph. It is still easily observable how the network has already established connections through many central nodes. An example of a MultiDiGraph, commonly referred to as a labeled directed graph in mathematics, is however shown in Figure 5.2.

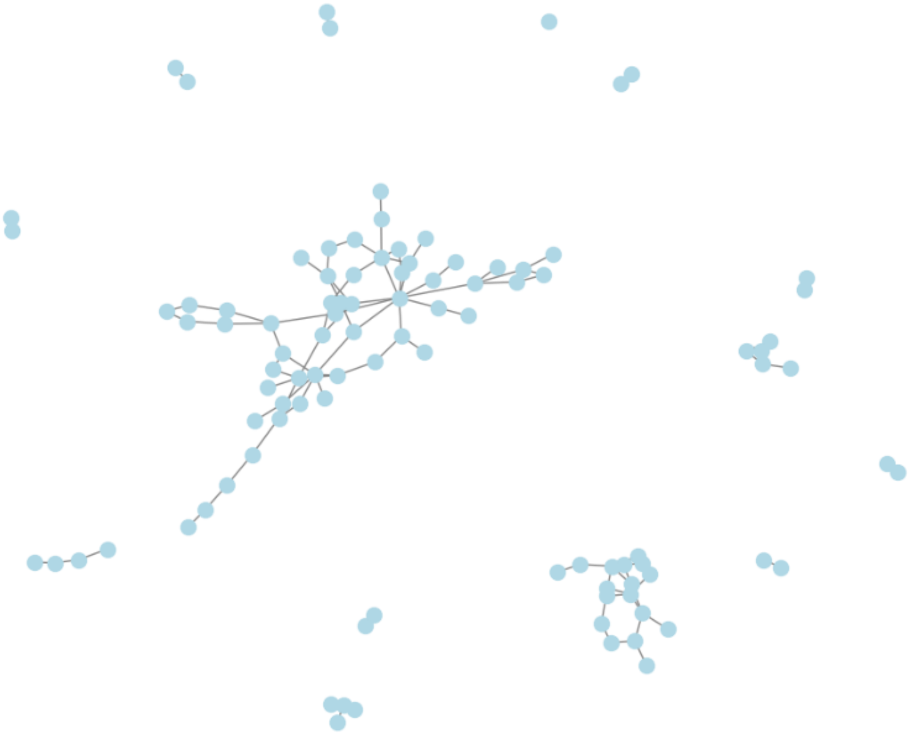


Figure 5.1: Subgraph of the first 100 nodes in the graph.

[2] Figure made using draw.io

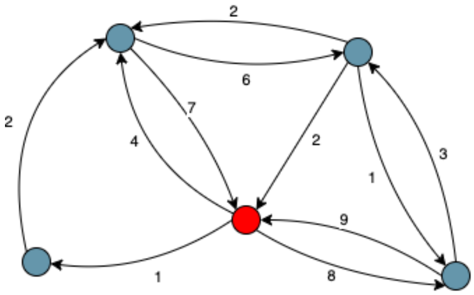


Figure 5.2: Example of a NetworkX MultiDiGraph[2]

5.2 Clustering Coefficient

The MultiDiGraphs allowed us to extract essential information concerning normal users and UoIs. We learned earlier from the work of [Eng23a] that the CC was an interesting feature to explore, which is explained in Section 2.3.2. NetworkX includes an algorithm called *clustering*, which computes CC for specific nodes. By utilizing this method and iterating it over the sets of normal users and UoIs, we were able to calculate the average CC for both groups.

The application of this method resulted in an average CC for normal users of **0.063** and **0.019** for the UoIs. This was an important finding, as the normal users CC is on average 69% higher. This confirms Eng's [Eng23a] statement, and gives us ground to use the CC as a feature to distinguish the UoIs from the normal users.

5.2.1 Other centrality measures

We also explored the idea of comparing several different centrality metrics, such as betweenness-centrality, closeness-centrality, and eigenvector centrality [Ruh00]. However, these were dismissed because of lack of relevance to our problem, and most of these were concluded to be too computationally demanding as well.

5.2.2 Degree measures

The MultiDiGraph also lets us calculate other node-specific features related to the weight and direction of the edges. Earlier in this chapter, we discussed the neighbor counts for the user types, which correspond to the node degree. Using the NetworkX methods *in-degree* and *out-degree*, we were able to calculate a variety of features related to the node degrees that we thought could differentiate the two user groups. Specifically, we computed the average *Ratio of Weighted In Degree* and *Ratio of Weighted Out Degree*, as described by the following formulas:

The *Average Ratio of Weighted Out Degree* is given by:

$$\begin{aligned}&\text{Average Ratio of Weighted Out Degree}\\&=\frac{1}{|S|}\sum_{n\in S}\left(\frac{\deg_w^-(n)}{\deg_w(n)}\right)\end{aligned}\tag{5.1}$$

where $\deg_w^-(n)$ is the weighted in-degree of node n (the sum of weights of incoming edges), and $\deg_w(n)$ is the total weighted degree of node n (the sum of weights of all edges connected to n).

Similarly, the *Average Ratio of Weighted In Degree* is given by:

$$\begin{aligned}&\text{Average Ratio of Weighted In Degree}\\&=\frac{1}{|S|}\sum_{n\in S}\left(\frac{\deg_w^+(n)}{\deg_w(n)}\right)\end{aligned}\tag{5.2}$$

where $\deg_w^+(n)$ is the weighted out-degree of node n (the sum of weights of outgoing edges), and $\deg_w(n)$ is again the total weighted degree of node n.

These ratios provide insight into the balance between incoming and outgoing interactions for users, allowing us to assess whether the activity patterns differ between normal users UoIs.

The results are depicted in Table 5.1, which shows that the UoIs has a higher average in-degree and out-degree. This is to be expected, as the UoIs on average interact with more users. What we were actually interested in exploring, was if there was a significant difference between the in-degree and out-degree between the two user types. The normal users had an very small difference of 0.41%, while the UoIs had a much higher difference of 4.09%. We did expect a more substantial difference for the UoIs, but this still confirms that they are more likely to be sending messages rather than receiving them.

	Normal users	**UoIs**
Average in-degree	1902.55	2717.02
Average out-degree	1910.37	2828.17
Difference (%)	0.41%	4.09%

Table 5.1: Average in and out degree of users.

5.3 Conversation Initiation Rate (CIR)

As "normal behavior" in online platforms is to exchange messages with people you know, we assumed that UoIs build their network by initiating conversations with strangers. Hence, it was interesting to study how often the users in the user groups initiate a conversation compared to how often someone initiates a new conversation with them. In studying our network graph, we found that we could not directly measure how often users started conversations. To get around this, we went back to the original dataset and first isolated all the different conversations that our normal users and UoIs were part of. Then we checked to see if each user was the `initiator` or `receiver` of the first message in each context. This resulted in what we have called CIR. The results are shown in table 5.2, and were calculated equation 5.3.

$$\text{CIR} = \frac{\text{Number of Conversations Initiated}}{\text{Total Unique Conversations}} \tag{5.3}$$

User group	**Average CIR**
normal users	0.577
UoIs	0.783

Table 5.2: Average CIR for the users.

We anticipated that the CIR for normal users would be approximately 0.5. However, the observed CIR was slightly higher at 0.577 in our primary dataset. This deviation is likely related to the initial selection criteria of the dataset, which favored contexts that exceeded five messages in length. Despite this, our main interest lies in examining the differences between the two user groups, which is approximated to 35.70%. Thus, the CIR stands out as an important feature to include in our system.

Furthermore, we wanted to see how often the user sends one message to initiate a conversation without receiving a reply, which was also pointed out by Eng [Eng23a]. By using equation 5.4, we discovered that the difference was less than 1%, making this feature negligible in our system. In this formula, $|S|$ represents the total number of users in the set S, $\deg_1^+(n)$ denotes the number of times user n sends a single message to initiate a conversation (assuming the user would send over 1 message if an reply happened), and $\deg_{\text{out}}(n)$ is the total number of outgoing messages sent by user n.

$$\text{Single Message Initiation Ratio} = \frac{1}{|S|} \sum_{n \in S} \left(\frac{\deg_1^+(n)}{\deg_{\text{out}}(n)} \right) \tag{5.4}$$

User group	**Single Message Out Ratio**
Normal users	0.324
UoIs	0.320

Table 5.3: Average Single Message Out Ratio.

We determined that the difference in the Single Message Out Ratio between the two user groups was too little to serve as a reliable distinguishing factor.

5.4 Choice of Features: Summary

Upon extracting a number of features from the dataset, we have identified a subset of features that effectively differentiate UoIs from normal users. These features are selected not just for their ability to distinguish the user groups, but also for their computational efficiency, as they are likely to be implemented in real-time ranking system. Initial findings indicate that normal users typically exhibit a lower neighbor count, compared to UoIs, which are associated with a higher amount neighbor. This particular metric can be easily calculated using the `degree` function provided by the NetworkX library. Furthermore, our analysis revealed a trend in which UoIs hold a significantly lower CC, which is also easily calculated with the `clustering` method applied to the user nodes. Another notable distinction is that UoIs generally exhibit a higher CIR. These features are node-specific, meaning that they are unique for each user. The 3 node-specific features we chose to proceed with when designing the system are shown in Table 5.4.

Feature	**normal users**	**UoIs**	**Difference**
Neighbors	152.19	224.61	47.58%
CC	0.06258	0.0193	-69.19%
CIR	0.577	0.783	35.70%

Table 5.4: Comparison of normal users and UoIs features

Additionally, we have incorporated a context-specific feature, documented by Eng [Eng23a], which states that UoIs engage more frequently in conversations that span five messages or fewer. This will be taken into account when designing the system.

Chapter 6
System design

Our next challenge consisted of how we would implement the features identified in Chapter 5 to create a ranking system, ensuring that its computational efficiency is sufficient to run in real time. We decided to base our system on the principles of fuzzy logic, as explained in Chapter 2.4. More specifically, we were drawn to the capability of fuzzy logic to compute "degrees of truth", providing a nuanced alternative to the binary "true or false" in traditional Boolean logic. These degrees of truth will in our case correlate with the increase or decrease in risk after an interaction with a user, which we will call a "risk change". This will become clearer after a more in-depth description of how we constructed the FDT, explained in this chapter.

6.1 Conditional Statements in the FDT

The conditional statements that the FDT is built upon are shown in table 6.1.

Conditional Statement	Evaluation Criteria
Receive message	Action is "receive" or "send"
Is this a new neighbor	Is the interacting user in the user's ego-graph
Conversation length over 5	Is the total conversation weight over 5
CC over threshold	True/False
CIR over threshold	True/False
Neighbor count over threshold	True/False

Table 6.1: Overview of Conditional Statements Used in the FDT.

The main idea of the FDT is to map every direct interaction with a user to an endpoint, involving the features identified in Chapter 5. For every interaction, the user iterates through a range of conditional statements, eventually ending up at an endpoint. After iterating through every interaction of a user, we can see how the interactions are distributed among the endpoints. The objective of this procedure is to run the script over a selection of normal users and UoIs, aggregate their respective scores, and then display the different distributions. This allows us to identify endpoints that are characteristic of normal users and UoIs, and therefore provides a basis to later assign risk changes to each endpoint. This process constitutes the training phase of our model, where we learn and establish patterns in user behavior that are indicative of potential predatory behavior. Technically, the FDT operates as a standard decision tree until a risk change is assigned to each endpoint. For simplicity, however, we will continue to refer to it as a FDT.

6.2 Construction of the FDT

The construction of the tree is rather straight forward. For each branch in the tree, a new feature/conditional statement is checked. Figure 6.1 shows the complete FDT. Due to its size, we have also divided the diagram into 3 parts in this document, each segment color-coded in Figure 6.1. Figure 6.2 depicts the left side of the FDT, handling instances when the user *receives* a message. Figures 6.3 and 6.4 show the right side of the FDT, which handles instances where the user *sends* a message. The diagram was created using diagrams.net [1].

One might notice that some endpoints do not go through all the levels or check of features, and there are numerous reasons for that. Firstly, Endpoint 1 in Figure 6.2 is already reached if the user receives a message from someone with whom they have never interacted before, that is, a new neighbor. We have isolated this specific incident as we found it interesting to explore whether this occurs at a higher rate for normal users compared to UoIs. In Figure 6.3, one can also observe that the *"Total weight/number of messages over 5?"* check is skipped when there is a new neighbor. This is because a new neighbor naturally implies that there are fewer than 5 messages exchanged, allowing us to iterate the FDT at a slightly faster rate.

[1] The open-source diagram software diagrams.net.

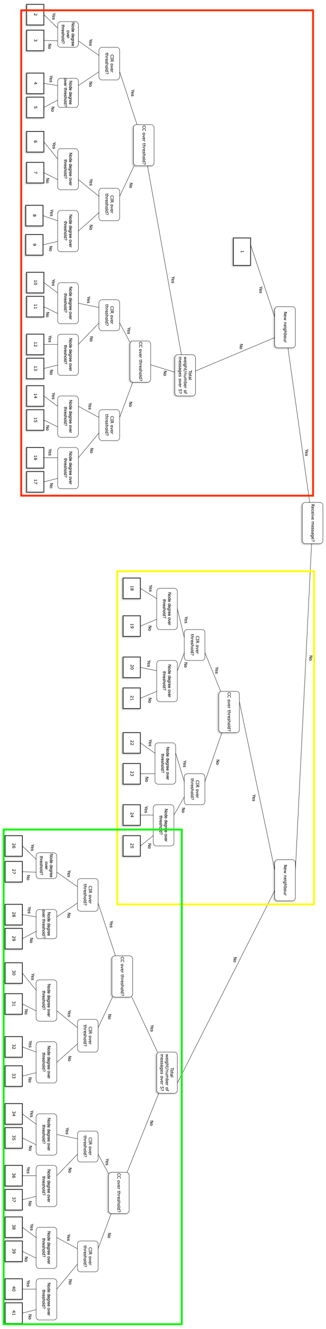


Figure 6.1: The figure shows the full tree, where the coloured rectangles displays the zoomed in versions of the figure. Red is Figure 6.2, yellow is Figure 6.3 and green is Figure 6.4.

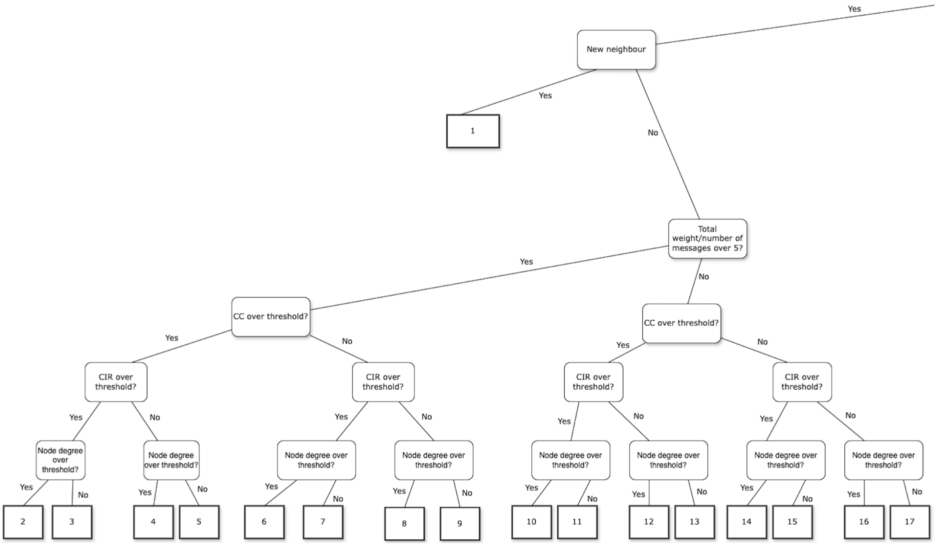


Figure 6.2: The FDT's left side for received messages, marked in red in Figure 6.1.

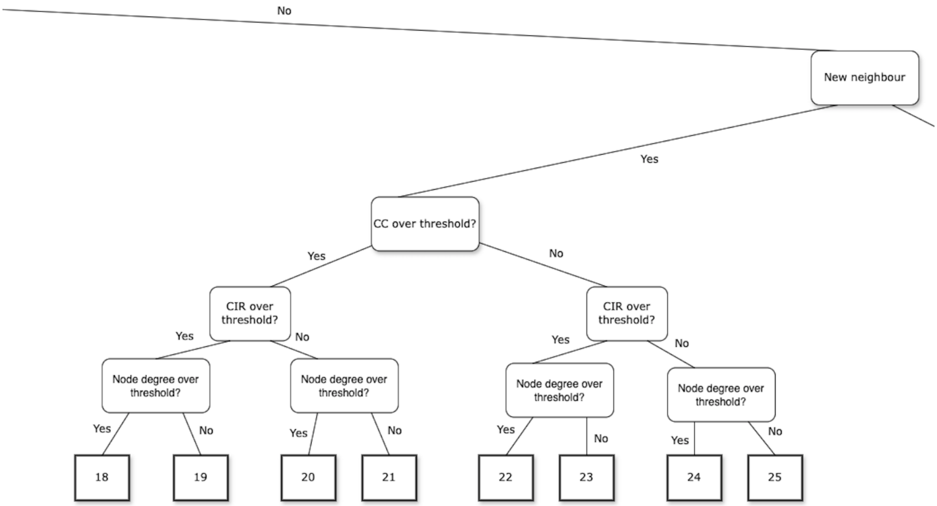


Figure 6.3: The FDT's right side for received messages, marked in yellow in Figure 6.1

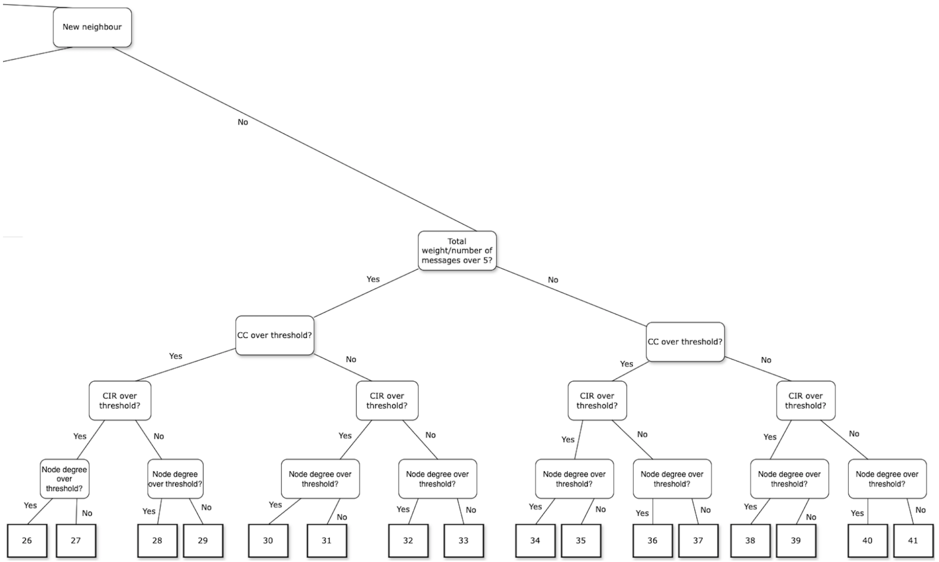

New neighbour
No
Total weight/number of messages over 5?
Yes
No
CC over threshold?
CC over threshold?
Yes
No
Yes
No
CIR over threshold?
CIR over threshold?
CIR over threshold?
CIR over threshold?
Yes
No
Yes
No
Yes
No
Yes
No
Node degree over threshold?
Node degree over threshold?
Node degree over threshold?
Node degree over threshold?
Node degree over threshold?
Node degree over threshold?
Node degree over threshold?
Node degree over threshold?
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
Yes
No
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41

Figure 6.4: The figure shows the FDT's right side for received messages, marked in green in Figure 6.1.

Feature	Threshold
Total weight	5
CC	0.042
CIR	0.63
Node degree	180

Table 6.2: Features and corresponding thresholds

In the tree, the thresholds we initially chose were in the middle range between the average values of the normal users and the UoIs, previously acquired in Table 5.4. However, CIR is a special case here, where the respective values for normal users and UoIs were 0.57 and 0.78. The normal users' values were artificially high due to the selection criteria of the dataset, mentioned in Chapter **??**. As a result, we opted for a somewhat lower threshold for the CIR than the middle point to compensate for this anomaly.

6.2.1 Code implementation

The code defines a Python function called `process_interactions(df, user)` that takes a DataFrame `df` and a specific user identifier `user`.

The function starts by initializing a new Graph[2], represented using the NetworkX library, with the user as the initial node. We have used a simple undirected graph rather than a MultidiGraph, as we do not need to know the weights in different directions to calculate the total weight. The function then identifies interactions directly involving the user from the DataFrame, constructs a set of neighbors based on these interactions, and also filters the DataFrame for relevant interactions between neighbors.

Next, the function sorts all interactions chronologically and iterates through each, updating the graph with nodes and edges representing the relationships. If an interaction involves the user, the graph is updated with an appropriate weight to reflect the number of messages exchanged between this user and the neighbor. The function then determines whether the interaction is a message sent or received by the user and applies the FDT logic through separate functions, `decision_tree_on_receive` and `decision_tree_on_send`. These two functions include the functionality to go through the whole FDT, including checking the features mentioned in Table 6.2.

If the interaction is solely between neighbors of the user, it ensures that an edge exists in the graph. This is to ensure that the CC is calculated correctly, as it quantifies the relationship between neighbors. The `process_interactions` function finally returns the endpoint counters, which will be used for further analysis.

6.2.2 Action distribution

As mentioned in the previous section, the FDT maps each user interaction to an endpoint using the features outlined in Chapter 5. This process, when applied to a number of normal users and users of interest (UoIs), reveals the distribution of endpoints. This helps us understand which endpoints are more commonly associated with UoIs compared to normal users. Consequently, it provides a foundation for assigning a risk change to the specific endpoints. After 'x' interactions, a user will have an aggregated *risk score*, which combines all the

[2] NetworkX documentation for the graph type "Graph"

individual risk changes from each interaction. The risk score is a measure designed to identify users whose interactions exhibit patterns indicative of predatory behavior. Importantly, this score will not fall below zero, as we are only interested in the UoIs.

To perform this analysis, we divided our users into two sets: a "training set" and a "test set." The list of UoIs contains a total of 58 users, while the list of normal users comprises more than 1,800 individuals. For the training set, we selected 2/3 of the UoIs, alongside an equivalent number of normal users. This approach allowed us to learn the distinguishing features and endpoints of each user group. The remaining 1/3 of the UoIs formed the "test set," which will be used later to evaluate the effectiveness of our FDT. The result of the initial train phase is shown in Table 6.3.

Table 6.3: Results of of UoIs and normal users interactions by endpoint. Mark by color instead of bold later.

Endpoint	Users of Interest	Normal Users	Difference
1	0.76%	1.38%	-0.62
2	0.00%	0.00%	0.00
3	2.07%	0.60%	1.47
4	0.00%	0.40%	-0.40
5	1.41%	14.34%	**-12.93**
6	11.15%	4.69%	6.46
7	20.65%	6.01%	**14.64**
8	0.00%	5.02%	-5.02
9	5.67%	14.11%	-8.44
10	0.00%	0.00%	0.00
11	0.27%	0.17%	0.10
12	0.00%	0.08%	-0.08
13	0.07%	0.79%	-0.72
14	2.39%	0.50%	1.89
15	3.01%	0.88%	2.13
16	0.00%	0.44%	-0.44

Continued on next page

Endpoint	**Users of Interest**	**Normal Users**	**Percentage Point Difference**
17	0.48%	1.18%	-0.70
18	0.00%	0.00%	0.00
19	0.15%	0.10%	0.05
20	0.00%	0.03%	-0.03
21	0.01%	0.22%	-0.21
22	1.62%	0.22%	1.40
23	2.01%	0.48%	1.53
24	0.00%	0.16%	-0.16
25	0.15%	0.29%	-0.14
26	0.00%	0.00%	0.00
27	1.87%	0.72%	1.15
28	0.00%	0.36%	-0.36
29	1.25%	15.00%	**-13.75**
30	11.14%	4.06%	7.08
31	22.24%	5.81%	**16.43**
32	0.00%	4.75%	-4.75
33	5.65%	13.26%	-7.61
34	0.00%	0.00%	0.00
35	0.22%	0.16%	0.06
36	0.00%	0.07%	-0.07
37	0.07%	0.83%	-0.76
38	2.22%	0.44%	1.78
39	2.90%	0.84%	2.06
40	0.00%	0.39%	-0.39
41	0.54%	1.22%	-0.68

The data presented in the *difference* column provides clear evidence that the FDT is able to verify differences in interactions and related features between the two user groups. For example, there are four endpoints where the percentage point difference exceeds ±10. These are endpoints 5, 7, 29, and 31, where their respective paths are shown in Tables 6.4, 6.5, 6.6 and 6.7. The endpoints with a high negative difference can be deemed "good", while the endpoints with a high positive difference can be deemed "bad".

Endpoint 5	
Feature	**Condition**
Receive message	True
New neighbor	False
Total weight over 5	True
CC over threshold	True
CIR over threshold	False
Node degree over threshold	False

Table 6.4: Path to Endpoint 5, which is a "good" endpoint"

Endpoint 7	
Feature	**Condition**
Receive message	True
New neighbor	False
Total weight over 5	True
CC over threshold	False
CIR over threshold	True
Node degree over threshold	False

Table 6.5: Path to Endpoint 7, which is a "bad" endpoint"

From the tables, we can identify two features that differentiate the two endpoint pairs. We call them pairs because two of them involve the user receiving a message, while the other two involve the user sending a message. There are several combinations of features that differentiate the users; however, for purposes of this analysis, we select the most polarized examples. The distinguishing features include "CC over threshold" and "CIR over threshold." UoIs are below the CC threshold but above the CIR threshold, whereas the normal users exhibit the opposite pattern.

Endpoint 29

Feature	**Condition**
Receive message	False
New neighbor	False
Total weight over 5	True
CC over threshold	True
CIR over threshold	False
Node degree over threshold	False

Table 6.6: Path to Endpoint 29, which is a "good" endpoint"

Endpoint 31

Feature	**Condition**
Receive message	False
New neighbor	False
Total weight over 5	True
CC over threshold	False
CIR over threshold	True
Node degree over threshold	False

Table 6.7: Path to Endpoint 31, which is a "bad" endpoint"

6.2.3 Risk score assignment

Based on the differences in Table 6.3, we were able to assign risk changes to the endpoints. We decided to include only those endpoints where the percentage point difference exceeded 2, where we have assigned risk scores ranging from -1 to 3. The logic behind this approach is that if an individual engages in one "bad" behavior, the risk score increases by 3, requiring at least three "good" interactions, each decreasing the risk score by 1, to counterbalance it. Thus, a higher maximum increase is intentionally assigned compared to the maximum decrease of the risk. In regular conversations, a user who occasionally engages in "bad" interactions that increase their risk score, will typically balance this out with subsequent positive interactions, due to the generally normal and varied nature of their communication pattern. These positive interactions will gradually decrease the risk score, as they are more reflective of the user's usual behavior.

Conversely, in predatory conversations, the individual with malicious intent is less likely to suddenly change to a "good" behavioral patterns. This reduces the likelihood of engaging in normal, risk-reducing interactions. Thus, the risk score stays high as predatory behavior isn't balanced by benign exchanges. As a result, in theory, the system is better able to distinguish between genuinely risky users and those with sporadic negative interactions within the context of predominantly normal behavior.

The risk scores for each endpoint are shown in Table 6.8, where endpoints with a risk change of 0 are excluded.

Table 6.8: Results of UoIs and normal users interactions by endpoint *with* risk scores.

Endpoint	**Users of Interest**	**Normal Users**	**Difference**	**Risk Change**
5	1.41%	14.34%	-12.93	-1
6	11.15%	4.69%	6.46	1
7	20.65%	6.01%	14.64	2.5
8	0.00%	5.02%	-5.02	-0.45
9	5.67%	14.11%	-8.44	-0.75
15	3.01%	0.88%	2.13	0.1
29	1.25%	15.00%	-13.75	-1
30	11.14%	4.06%	7.08	1.5
31	22.24%	5.81%	16.43	3
32	0.00%	4.75%	-4.75	-0.4
33	5.65%	13.26%	-7.61	-0.6
39	2.90%	0.84%	2.06	0.1

Chapter 7

Testing and Results of Test Data

This chapter presents the results and subsequent discussions of the different variations of the system. First it covers two different variations of the system, where we used the remaining 1/3 the UoIs in the initial dataset for testing. Section 7.1 covers the initial system, with the related results and analysis. The analysis is the foundation for the improvements covered in Section 7.2. In this section, a new test with the same dataset is executed, with a brief analysis of the results.

7.1 Initial System Testing

By mapping each endpoint to a corresponding risk change, we could simulate how the user risk changes over time. Instead of using a fixed time frame, we are using a fixed number of interactions, as it allows for a better comparison.

7.1.1 Results of the Initial System

In this test, we chose to have a baseline risk score of 50 to allow us to confirm that the user's risk score actually decreases by behaving "good". During this initial evaluation of the operational ranking system, the decision tree's feature thresholds remained the same as in the previous chapter (Table 6.2). The results are shown in Tables 7.1 and 7.2, which display the risk scores after 300 and 1000 interactions, respectively. Both tables are ordered by decreasing risk scores. Figure 7.1 depicts a selection of 4 UoIs and normal users risk score development over 300 interactions.

Table 7.1: Ranking of users after 300 interactions

Rank	User Type	UserID	Risk Score
1	User of Interest	238F	718.80
2	User of Interest	CCE0	706.10
3	User of Interest	207A	679.10
4	User of Interest	399D	653.20
5	User of Interest	AB7D	634.50
6	Normal User	087F	616.95
7	User of Interest	8821	607.90
8	User of Interest	7C8C	550.00
9	User of Interest	A1AB	470.50
10	User of Interest	CBFE	396.50
11	Normal User	066E	393.15
12	User of Interest	6307	283.05
13	User of Interest	F874	205.00
14	Normal User	0673	123.90
15	Normal User	095E	100.25
16	Normal User	0958	90.95
17	Normal User	08B1	57.25
18	User of Interest	1E0B	40.30
19	Normal User	084E	3.00
20	User of Interest	F83F	0.00
21	User of Interest	647A	0.00
22	Normal User	064E	0.00
23	Normal User	0636	0.00
24	Normal User	0633	0.00
25	User of Interest	E3D7	0.00
26	User of Interest	1625	0.00
27	User of Interest	10BA	0.00
28	User of Interest	7B18	0.00
29	Normal User	0696	0.00
30	Normal User	075D	0.00
31	Normal User	074C	0.00
32	Normal User	073F	0.00
33	Normal User	073E	0.00
34	Normal User	06DE	0.00
35	Normal User	0870	0.00
36	Normal User	08F2	0.00
37	Normal User	08BC	0.00

Table 7.2: Ranking of users after 1000 interactions

Rank	User Type	UserID	Risk Score
1	User of Interest	238F	2449.50
2	User of Interest	207A	2219.20
3	User of Interest	8821	2082.90
4	User of Interest	399D	2074.40
5	User of Interest	CCE0	2029.40
6	Normal User	087F	2017.90
7	User of Interest	7C8C	1761.20
8	User of Interest	10BA	1739.30
9	Normal User	066E	1705.35
10	User of Interest	AB7D	1692.90
11	User of Interest	A1AB	1574.90
12	User of Interest	CBFEC	478.80
13	User of Interest	6307	258.70
14	Normal User	08B1	257.45
15	User of Interest	F874	205.00
16	Normal User	0673	123.90
17	Normal User	095E	100.25
18	Normal User	0958	90.95
19	Normal User	084E	3.00
20	User of Interest	F83F	0.00
21	User of Interest	647A	0.00
22	Normal User	064E	0.00
23	Normal User	0636	0.00
24	Normal User	0633	0.00
25	User of Interest	1E0B	0.00
26	User of Interest	7B18	0.00
27	User of Interest	1625	0.00
28	User of Interest	E3D7	0.00
29	Normal User	0696	0.00
30	Normal User	075D	0.00
31	Normal User	074C	0.00
32	Normal User	073F	0.00
33	Normal User	073E	0.00
34	Normal User	06DE	0.00
35	Normal User	0870	0.00
36	Normal User	08F2	0.00
37	Normal User	08BC	0.00

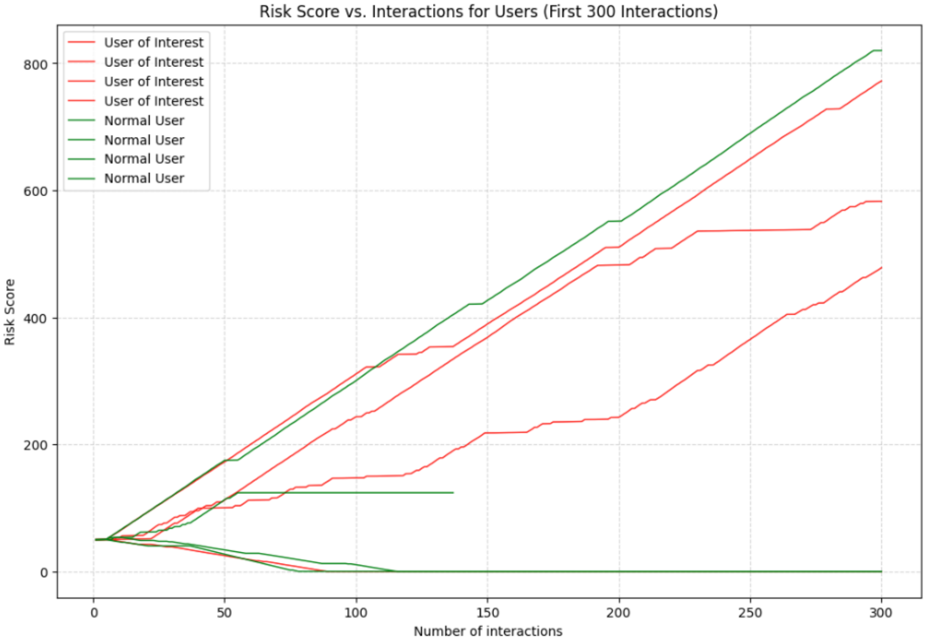


Figure 7.1: A selection of 4 UoIs and 4 normal users' ranking development over 300 interactions.

The especially high-risk scores are particularly significant, as they indicate users who may pose a greater threat or exhibit potentially harmful behavior. Some UoIs have notably high scores, including:

- User 238F with a risk score of 2449.50 (Rank 1)
- User 207A with a risk score of 2219.20 (Rank 2)
- User 8821 with a risk score of 2082.90 (Rank 3)

The presence of UoIs within the top ranks partially confirms the effectiveness of the risk scoring system in identifying individuals who could require closer moderation. From the tables, we can observe a trend that indicates that UoIs generally tend to achieve higher risk scores, However, there are more exceptions to this trend than initially anticipated. Although the majority of normal users receive a risk score of 0 or close to 0, there are numerous instances where normal users receive a higher risk score than expected. For example, at ranks 6 and 9, two normal users rank high, with another normal user at rank 14. An even more important concern is the amount of UoIs having a risk score of 0. The prevalence of zero scores among the UoIs suggests potential gaps in the system, which could indicate overly stringent thresholds or the need for more nuanced criteria.

An important consideration is that we have not closely examined the specific conversations in which these users are involved. This means that we have not reviewed all their conversations to comprehensively evaluate the likelihood of predatory behavior. During the initial labeling process, we identified users who initiated conversations of a sexual nature. However, some normal users might be "false negatives" in terms of labeling, meaning they are actually individuals we would classify as UoIs. This misclassification may have occured because we assumed that any user not explicitly labeled as a UoI was a normal user. Consequently, we randomly selected a sample of these presumed normal users for system testing, potentially including some misclassified UoIs.

7.1.2 Analysis of Initial System

In order to improve the system, we need to analyze what causes particular users to score the way they do. Our approach to do so involves examining the chats of every individual, as well as the values of features such as the CC, CIR, and neighbor count. We sat a maximum limit of 50 conversations per user that we could read through briefly, as some of the users have several hundred different conversations of various lengths. Many of the conversations were notably short, often consisting of five messages or fewer. This was one of the key reasons we decided to include a sample of 50 conversations in our review. By examining a larger number of interactions, we ensured that we captured a sufficient number of complete conversations to get a better understanding of the individual user's intentions.

From a quick overview, we can observe that there is a selection of users with a risk score above 2000. In order to obtain such a high score, they are frequently ending up in the the endpoints with the large risk changes. Those correspond to endpoint 6 and 7 when receiving messages and endpoints 30 and 31 when sending messages. The common factor among these are low CC and high CIR, both on the "bad" side of the threshold.

During this examination stage, we also wanted to try to separate sexters from users who are more likely to be predatory. A 2013 study [GBGZ13] provides evidence that sexting is common among young adults, but it appears to be not related to sexual risk or psychological health. However, distinguishing these is an inherently complex task because predators often go to great lengths to disguise themselves as individuals who are similarly aged to the children or teenagers they target. This deliberate mimicry makes it challenging to identify malicious intent. We therefore proceeded with an approach in which we marked some of the UoIs as high risk users. Some of the reasons for this were:

- The user is inconsistent when talking about their age.
- The user is very graphical in their language. For instance, one user talks alot about the size of both his own genitalia and the others user's size.
- The user immediately insists on sexual roleplaying or sexting.
- The user is clearly pushing on homosexual topics.

This resulted in a list of 13 high risk users with variable risk scores, which will be used when the system is further improved. We also noted the features values of the all users for further insight. One common factor between the high risk users is a very high CIR, averaging over 0.9. In addition, a very low CC, below 0.01, is commonly seen between these users. Also, user 087F (a normal user) was actually involved in several instances of sexting, and we therefore changed this user's label from normal user to UoI.

The top 3 users were userIDs 2384, 207A, and 8821. Based on the chats, user 2384 appears to be a female reaching out to many users. In many of the chats, the user initiates sexting early in the conversation. A complete analysis of all the user's interactions shows a CC of 0.033, CIR of 0.78 and neighbor count of 689, which are all over the initial thresholds. User 207A, who holds the second highest ranking, exhibits a CC of 0.013, an CC of 0.80, and a neighbor count of 554. From analyzing the chats, the user exhibits abnormal behavior through a lot of roleplay. The roleplay is not necessarily sexual, but is still physical through messages like *"*hugging u*"* and *"*starts kissing u*"*. The use of '**' indicates roleplay. These messages fromt the UoI often occur spontaneously and without prior context. At rank 3, user 8821 has a very low CC of 0.0055, CIR of 0.98 and a neighbor count of 404. This user also exhibits suspicious behavior, often asking abou count and initiating sexting frequently. Of these 3 users, those ranked second and third were marked as high-risk users.

One particular UoI is actually very likely to be a predator based on conversations. He calls himself "the therapist", and clearly indicates an interest for young girls through many of his conversations, where an example is shown in Table 7.3. What is particularly interesting about this user is his ranking and feature values. He has a risk score of 0 and "positive" feature values, which fall on the favorable side of the thresholds. His CC is 0.079 and CIR is 0.48. The endpoint he lands most frequently in is 5 when receiving messages, and endpoint 29 when sending messages, all of which have a maximum negative risk change (-1). These values categorize him as a normal user in terms of behavior based on the action distribution. Upon closer examination, it appears that girls initiate conversations with him as frequently as he initiates conversations with them. This reciprocal pattern of interaction suggests a mutual engagement in communication rather than a one-sided approach typical of the UoIs. These interactions can be quite graphic at times, and he tends to quickly suggest moving the conversation to other platforms. He also makes statements like "I want you to be younger" and does not seem to hide the fact that his behavior is highly predatory. A snippet of a conversation including this user is shown in Table 7.3.

Sender	**Content**
D5F0	hey
1E0B	yes?
D5F0	20 years later -_-
1E0B	ok
D5F0	u ok ?
1E0B	no
D5F0	what happen?
1E0B	girls
1E0B	2-15
D5F0	wha
1E0B	ages
D5F0	mine?
1E0B	stay on topic
D5F0	ok?
1E0B	well age
1E0B	yours

Table 7.3: Snippet of a conversation including a very hgh risk user.

During this exploratory stage, we also focused on normal users who achieved a high-risk score, as their results contradicted our expectations. Notably, the users ranked sixth and ninth drew our attention. The user in the sixth position was involved in several conversations involving sexting, suggesting that it could be labeled as a UoI. This user has a very low CC of 0.0087 and a high CIR of 0.86, which significantly contributed to its high ranking. The user ranked 9th is primarily engaged in begging for in-game items from various users. This behavior explains the high CIR value of 0.89 and the low CC value of 0.02, which contributes to her elevated rank. We can also find another instance of spamming at rank 14, where a user is promoting a discord server.

7.2 Refinement and Testing of the Improved System

This section covers how we modified the initial system and the results of those modifications.

7.2.1 Improving the Initial System

From the last section, we learned that the high risk users had a significantly higher CIR and CC than the threshold we sat initially. For the next test, we wanted to introduce more variations of the thresholds, rather than the binary above or below. This is the essence of fuzzy logic, as it allows for "degrees of truth". We therefore deemed it useful to introduce more extreme variations of the features, as it would provide a more nuanced scoring system. From the action distribution of the initial FDT, certain endpoints had very low degrees of utilization, meaning that users rarely hit them. This observation, in combination with the introduction of the extremes of the feature values, gave us the ground to create a more efficient design of the FDT. The new design is depicted in Figure 7.2, with the red rectangle showing Figure 7.3 and the green rectangle showing Figure 7.3b.

One significant modification was the removal of the "total weight" feature, as its influence on the ranking was deemed negligible. For the thresholds, we introduced extra variations through additional branches, thereby expanding the tree, increasing for 2 options from two to 3 options. The following changes were made:

- For the **CIR**, we introduced three intervals: below 0.63, between 0.63 and 0.90, and above 0.90 (extreme).

- For the **CC**, we also introduced three intervals: below 0.01 (extreme), between 0.001 and 0.042, and above 0.042.

These values were based on the feature values of the high risk users, who tended to have more extreme values. The new construction of the decision tree meant that we also needed to modify the risk changes related to the endpoints. We wanted to assign very high risk score changes to the "extremes", users with very low CC and very high CIR, and a smaller increase to the users with above average (but not over the extreme) feature values. We also changed the interval of risk change from -1,3 to -2,3, to further be able to reward "good" behavior. The risk change related to every endpoint can be seen in Table A.2 in Appendix A.

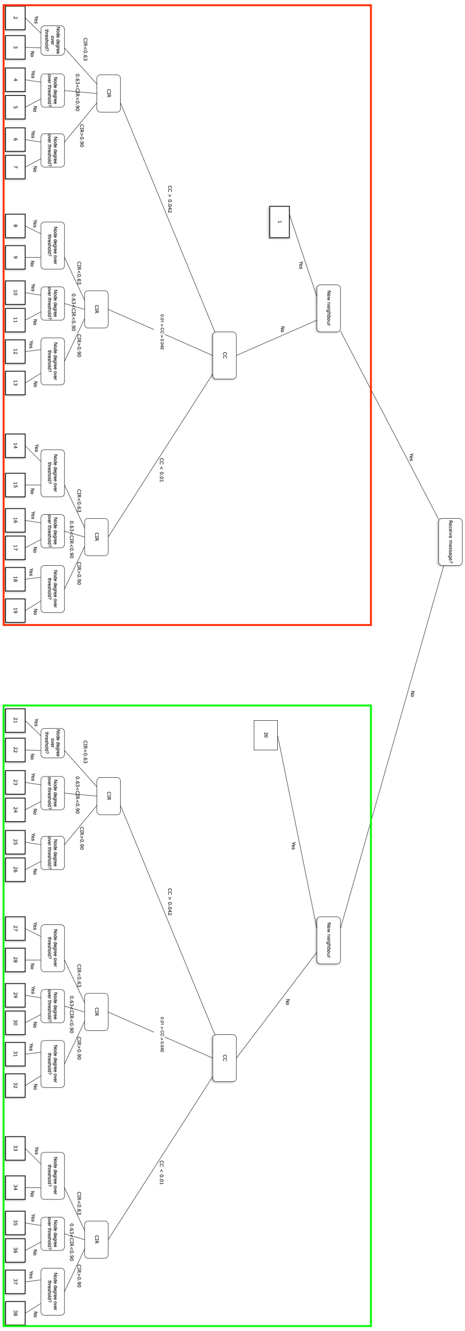
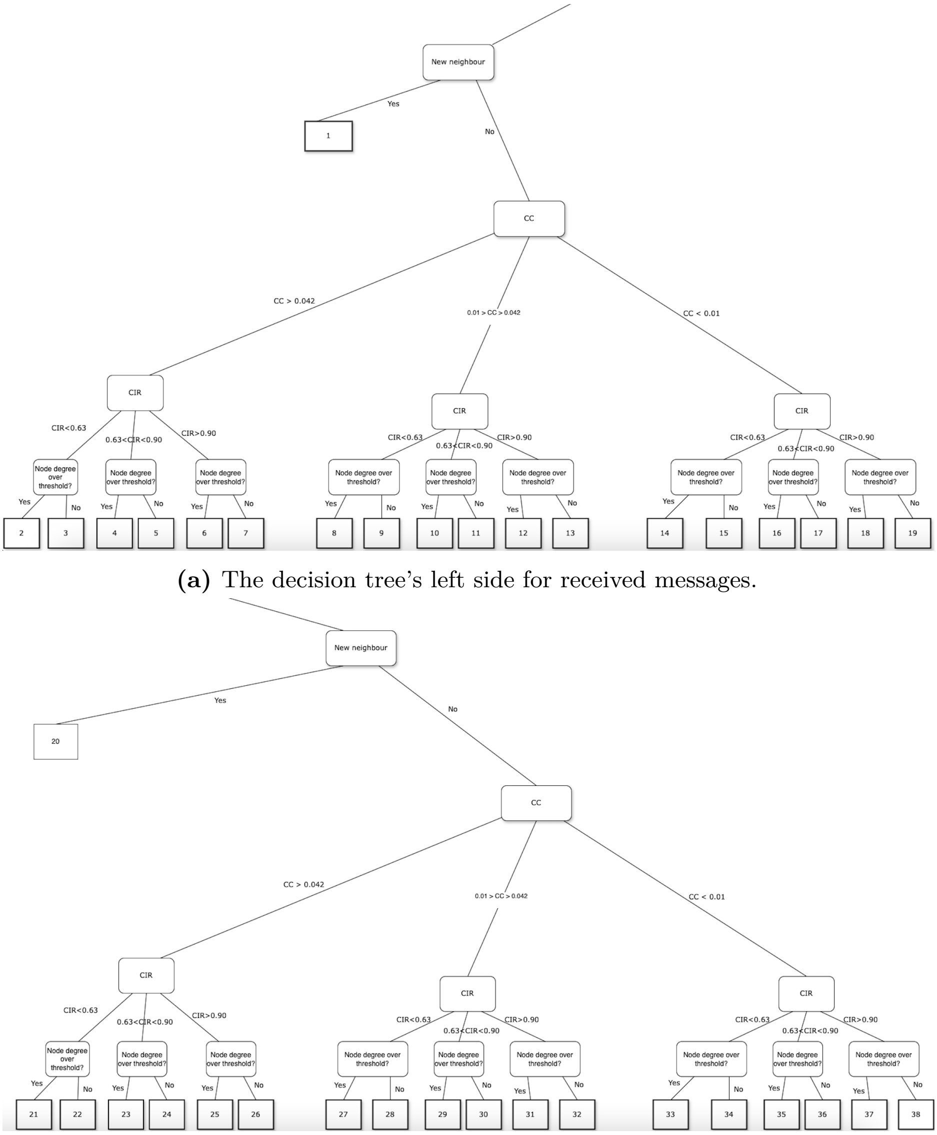


Figure 7.2: The figure shows the full design of tree 2, with the red rectangle showing Figure 7.3 and the green rectangle showing Figure 7.3



(a) The decision tree's left side for received messages.

(b) The decision tree's right side for sent messages.

Figure 7.3: The figures show the decision tree's sides for received and sent messages.

7.2.2 Results of Improved System

Table A.1 in Appendix A shows the results after a maximum of 1000 interactions on the same dataset as the initial system was tested on.

The results show a trend where UoIs and high-risk users advance in ranking, while most normal users drop in rank. Given that the primary dataset consisted of only 58 users, 1/3 of those being used for testing, we believe that the system's performance would improve with a larger and more diverse user base. A broader dataset would provide a wider range of interaction patterns and behaviors. Another concern regarding testing on the selected dataset is that not all users reach the maximum interaction limit. That means, a high risk user might have a lower rank than a normal user due to the fact that it has fewer interactions or that its feature values has not stabilised yet. Both these concerns will be addressed in the final system test in the next chapter.

Chapter 8
Final System Testing

This chapter covers the result and discussion of the final system test, where we test the improved system on a different *unlabeled* dataset containing 500 users.

8.1 Data Selection for Final Test

For the final test of the system, we selected 500 random users from the entire dataset of 214,848 users based on the criterion that they had engaged in more than 500 interactions in total. This approach was chosen to ensure that no users received artificially low scores due to insufficient interaction data, thereby providing a more accurate assessment of the system's effectiveness. This method helps in mitigating any biases that might arise from users with limited interaction history. In previous tests, this criterion was not met because we did not consider the number of interactions during the selection of the initial dataset, described in Chapter 4.2.

The system used, including thresholds and endpoint risk changes, was the same as in Section 7.2. The results were extracted into a CSV file containing the userID and risk score, along with the features CC, CIR and neighbor count. The users in this were unlabeled and we did not have other information except content of the csv file to decide whether they were a normal user or a UoI. We therefore had to manually examine the top-scoring users' chat history in order to assess what type of users received the highest scores. This was essential to evaluate the performance of the system.

8.2 User Examination

To add contexts to the the ranking after running it over 500 interactions, we needed to examine the top 50 ranked users first. Below is a list of some of the reasons a user was marked as a UoI, often also being a combination of these factors.

- **Sexting**: The user engages in sexual roleplaying.
- **Sexual Explicit Language:** The user use explicit language, such as saying they are "hard" or "laying in bed".
- **Immediate Sexual Requests:** The user has sexual enquires to the other user, such as asking for the number of sexual partners and masturbation habits.
- **Social Media Requests:** The user often pushes to get the social media contact info quickly into the conversation.
- **Ignoring Age Concerns:** The user does not care about the age of their conversation partner, persistently pushing for sexual interactions even when the partner mentions being underage.
- **Inconsistent Age :** The user changes its age based on the conversation, trying to relate more to responding part.

Figure A.1 and Figure A.2 in Appendix A provides user risk scores, CCs, neighbor count, CIR, and specific notes high scoring and low scoring usera after 500 interactions. The notes highlight the reasons for their labeling as good or bad, and provide context for their chat behavior. We also marked every user with one of four colors based on our interpretation of their chat history.

- **Red**: UoI or in other words, a bad user.
- **Green**: a good user.
- **Yellow**: non-sexual, high-scoring users exhibiting unwanted behavior.
- **Blue**: users placed artificially high due to rare feature combination or other.

8.3 Final Results

The distribution of the user categories (colors) is shown in Table 8.1. Figure 8.2 displays the distribution of risk scores among the 500 users, showing that the most common risk score is 0.

Color	Count	Percentage (%)
Red	28	56.0
Yellow	5	10.0
Blue	4	8.0
Green	13	26.0

Table 8.1: Color distribution among top 50 users.

If you disregard the blue user placed artificially high, red UoIs take up 60% of the distribution among the top 50 users. Figure 8.1 illustrates the distribution of colors among these top users, highlighting a clear trend where UoIs contribute significantly to the high density of red at the top positions. This indicates that the system effectively identifies and ranks high-risk users prominently.

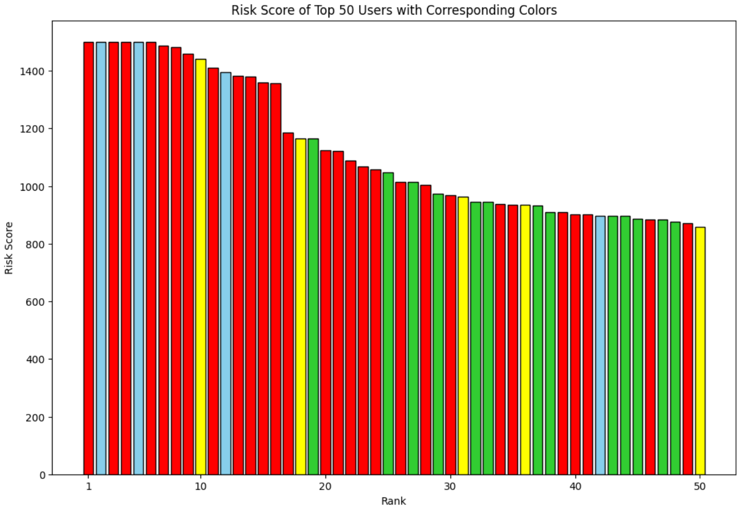


Figure 8.1: The figure captures the density of the colors, where the distribution of UoIs (red) is higher among the top ranks.

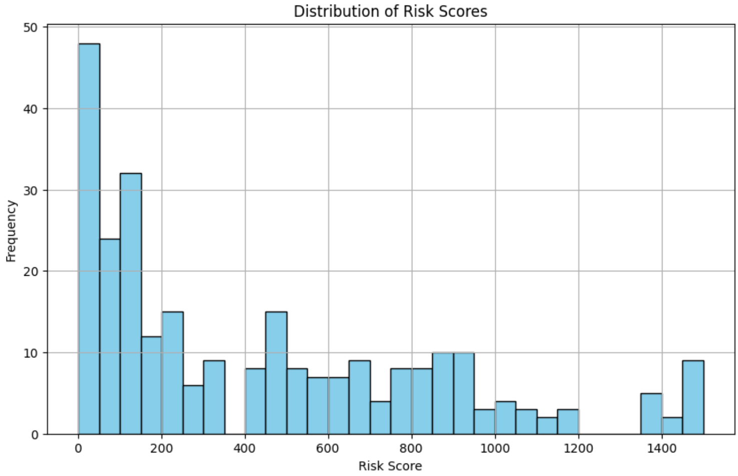


Figure 8.2: Risk distribution among all users.

8.3.1 Analysis of Final Results

Among the top 20 users, 14 of them were marked as bad. This was due to many reasons, the most common being sexting. Interestingly, all users in the top six ranking had a CIR of 1.00, resulting in a maximum risk score of 1500 because they automatically landed in a +3 risk score endpoint (endpoints 19,20 or 38). Four of the users with a CIR of 1.00 had a CC of 0.00 due to having only two isolated neighbors, while one user had 50 neighbors, resulting in a non-zero CC value.

Despite the potential irrelevance of the four users with only two neighbors, we decided to analyze their chat histories. Among these four, two were identified as UoI, one was a normal user, and one was a French-speaking user that we did not examine due to the language barrier. The presence of UoIs among those with two neighbors and rare feature values might be coincidental, but we decided to include them in the analysis regardless. An ego graph corresponding to users with a CIR of 1.00 and CC of 0.00 is illustrated in Figure 8.3, with the red node being the high scoring user and the red edges indicating the initiating connection.

To further evaluate the system's performance, we examined the chat histories of 50 users with the *lowest* risk scores. We also included a few users with risk scores slightly above 0 for diversity. Among these users, only two were identified

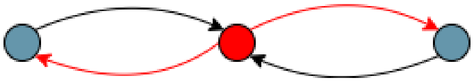


Figure 8.3: Ego graphs of users with CIR of 1.00 and CC of 0.00, with the red edge indicating the initiating connection.

as UoIs, while the rest were classified as "good". This part of the ranking can be seen in Figure A.2.

Notably, one of the UoIs had a CC of 0.051 and a CIR of 0.42. A detailed review of this user's chat history revealed that he is a well-known sexual roleplayer, recommended by others within the community. This social endorsement explains the observed feature values, which indicate a moderately connected network with users reaching out to him.

8.4 Extended User Interaction Analysis

The presence of normal users among the top 50 risk scorers, as well as UoIs among the lowest scoring users, made us interested to see if this is a consequence of the relatively low number of interactions. Due to being in a late stage of the research, we did not have time to test this on all the users we examined earlier. To test this we chose four user from the top 50 scoring users, and 5 from the lowest scoring users. Among the four in the top 50 interval were two UoIs and two normal users, where the normal users were ranked higher than the UoIs after 500 interactions. Similarly, we chose four users from the bottom of the ranking with a risk score of zero. Among these, the two UoIs and two normal users.

This selection of users were simulated up to 1500 interactions. The results can be seen in Figure 8.4 and Figure 8.5, where the black dotted line illustrates the point where the last simulation stopped.

Our initial hypothesis suggested that the risk score of the UoIs would consistently rise in comparison to that of normal users, eventually surpassing their scores. Figure 8.4 confirms this, as both UoIs' graphs climb above those of the normal users. Notably, the UoIs maintain a steady rate of increase in their risk scores, while the normal users' rate of risk increase gradually slows down over time. Likewise, in Figure 8.5, the risk score of one UoI eventually starts to increase significantly after approximately 750 interactions.

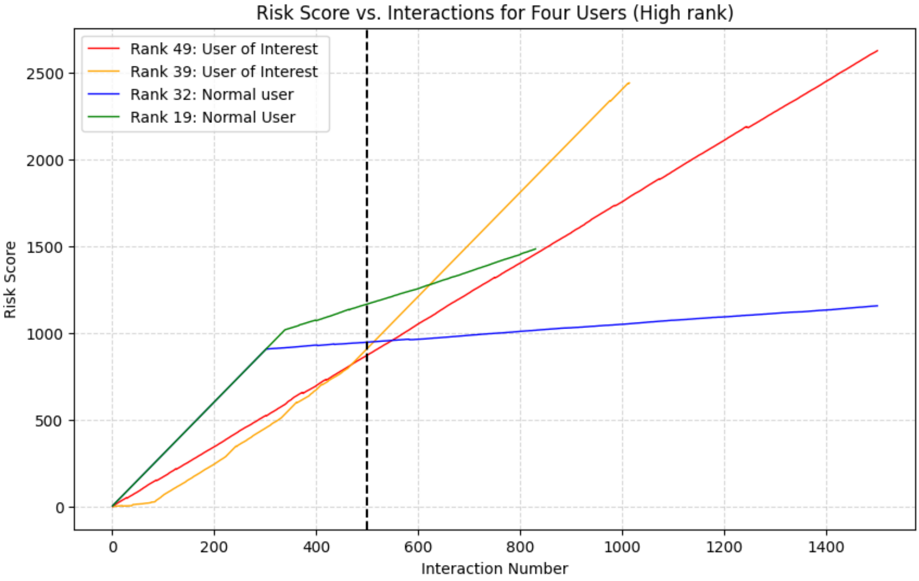


Figure 8.4: Development of 4 top ranked users beyond 500 interactions.

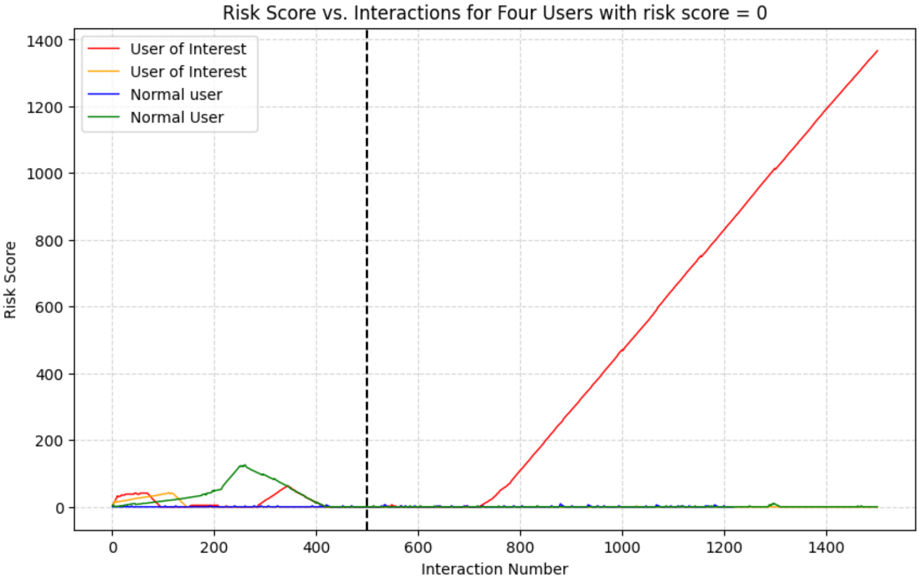


Figure 8.5: Development of 4 low ranked users beyond 500 interactions.

The normal users rate of risk increase gradually slowing down in Figure 8.4 is most likely due to the feature values finally "settling" to match their normal behavior. In the beginning, a low neighbor count and few interactions can lead to higher artificially high CC and CIR values that will eventually even out with normal interactions to more people. On the other hand, most UoIs will continue to show the same pattern of interactions.

In Figure 8.5, the other UoI maintains a very low risk score , which is due to the user's feature values not exceeding the thresholds. From the exploration in the previous section, we know that the user apparently is a recognized sexual role-player within the community, even being recommended by others. This explains why he keeps on ending in "good" endpoints with favourable negative risk changes.

Chapter 9
Discussion

This chapter provides a discussion of the methodologies, results, and implications of our systems, as well as debating its advantages and weaknesses.

9.1 Data Selection

Initially, 1000 random contexts with a length exceeding five messages were selected. This approach was partly driven by a collaboration with another group of researchers who focused primarily on conversation ranking rather than user ranking. By sharing the labeling duties, we aimed to achieve a more efficient workflow and enhance the overall effectiveness during this stage.

User-specific criteria, which was our main focus, were not considered in the selection of the initial dataset. This oversight may have resulted in many users within the dataset lacking an adequate number of interactions, thus not reaching the interaction limits set during testing. This will be explained further in the next section.

Although we do not have precise numbers on this, it is reasonable to assume that some UoIs in the systems discussed in Chapter 7.1 and Chapter 7.2 might have received higher risk scores if they had enough interactions reach the interaction limit. Conversely, this could have potentially lowered the scores of some normal users who were initially ranked high (further explained in Section 9.4).

Another important consideration is the number of UoI used for training and testing the model. Initially, we had a total of 58 UoIs from the first dataset. These were divided into two subsets, with approximately two-thirds allocated for training and one-third for testing. This distribution, while practical given the available data, suggests that our model's accuracy could benefit from a larger dataset. With additional time and effort dedicated to labeling a more extensive dataset, we could increase the number of UoIs, thereby enhancing the robustness of our training data. A larger sample size would provide a more reliable foundation for establishing feature threshold values, potentially leading to more precise and accurate identification of predatory behavior. This expansion would not only improve the model's training but also offer a more comprehensive evaluation during the testing phase, ultimately resulting in better performance and reliability.

In Chapter 8.1, which covered the final test and results, a new dataset was selected, with 500 users with over 500 interactions being the selection criteria. This ensured that the ranking was made on a similar basis for every user.

9.2 Feature Selection

In our analysis, the CC and CIR emerged as the most impactful features for identifying high-risk users. The CC provided a clear measure of the inter-connectivity within a user's network, with low CC values often correlating with predatory behaviors such as sexual roleplaying and inappropriate interactions. The introduction of extreme variations of the CC thresholds proved to be efficient, as the most high risk users tended to have values below 0.01.

Similarly, the CIR was crucial in identifying how often users initiated conversations, where UoIs tended to reach out to strangers rather than being approached themselves. We believe that the introduction of extremes here as well contributed to a more accurate system.

In contrast, the weight of the conversation was not as impactful in our system as expected. This was discovered during the action distribution, where it rarely distinguished UoIs and normal users. However, we still believe that it could be useful in combination with a neighbor count. A user with a high degree of low weighted conversations, in combination with a high neighbor count, could be an indication of predatory behavior.

9.3 FDT Construction

The construction of the FDT plays a crucial role in the effectiveness and accuracy of our system. Despite already incorporating numerous variations of features through additional branches, we believe that further expanding the tree is possible without significantly increasing computational complexity. Essentially, the FDT acts as a series of nested if-statements, meaning that adding more variations can be done efficiently. Expanding the tree involves identifying and integrating new feature variations that can capture subtle differences in user behavior.

The current system effectively moves UoIs towards the top of the ranking based on their risk scores. However, by integrating more nuanced feature variations into the FDT, the system could further refine this ranking. Adding these detailed features would enhance the granularity of the decision-making process, enabling the tree to better differentiate between varying levels of risk among users, especially UoIs. This could help in accurately identifying the most dangerous users and placing them at the very top of the ranking, ensuring that the system gives those who pose the greatest threat the highest risk score.

9.4 Interaction Limits

The final test was done with a limit of 500 interactions, making sure that every user reached this count. This limit resulted in a result where 56% of the UoIs landed in the top 50 highest scoring users. Among these 50 were other unwanted users, such as spammers, but also 26% normal users. We believe that the proportion of UoIs and other unwanted user would be even higher if we simulated with a higher number of interactions. As seen in Figure 8.4, UoIs tend to exhibit a steady increase in their risk scores due to their behavioral patterns consistently aligning with what we have identified as potentially predatory behavior.

In contrast, normal users may display brief periods where their behavior mimics predatory patterns, particularly in the early stages when their social networks are developing. A low neighbor count can initially skew results by producing a low CC and a high CIR. As the number of interactions increases, however, their feature values tend to normalize. This normalization ultimately places their interactions within "good" or at least outside the "extreme" endpoint. Consequently, their overall risk scores will ultimately decrease, or their risk score will increase less rapidly, which is seen in Figure 8.4 where we increase the interaction count from 500 to 1500.

To make an even more accurate system, a higher interaction count would provide a more comprehensive dataset, allowing for a better differentiation between normal users and UoIs. Increasing the number of interactions would help mitigate the initial skew caused by low neighbor counts, further normalizing feature values and enhancing the accuracy of the risk assessment. This improvement would make the system more robust in identifying predatory behavior and reducing false positives among normal users.

9.5 Key Advantages of Our System

As regulations and technologies continue to evolve to prioritize user privacy, our behavior-based system is well-positioned to adapt and remain effective. In the future, the ability to access and analyze the contents of user chats in *real-time* might be limited or entirely restricted. For many of the systems presented in previous studies, this presents a major challenge as they rely on content analysis. In such scenarios, a system like ours, which focuses on analyzing user behavior without examining the actual chat contents, could become particularly useful.

Our solution allows for dynamic and continuous risk assessment, enabling a live-ranking system that place users with the most predatory-like behavior among the top ranks. This allows the most dangerous cases to receive immediate attention, enabling human moderators to potentially intervene before harm is done. However, such intervention depends on human moderators having access to chat contents, which is necessary to verify and understand the context of the risk score. This represents a significant advantage over many previous studies which flags users once their risk levels exceeded a predefined threshold. In numerous of those cases, human moderators have to go through a list of flagged individuals without a clear understanding of which ones pose the greatest threat.

Scalability and flexibility are also important advantages. The use of a decision tree based on fuzzy logic allows our system to scale efficiently and process large volumes of interactions in real-time. It also allows for easy integration of additional features and modifications. This flexibility is essential for staying ahead of sophisticated and ever-changing predatory tactics, as well as evolving online environments.

9.6 Weaknesses and Challenges of Our System

In the online world, "normalizing" human behavior can be a big challenge. Internet culture evolves at a rapid pace, where smaller and bigger communities arise every single day. Individuals who might not have a large social circle in real life can find a sense of community through online interactions, such as playing simple games. The dynamic nature of online communities means that our system, based on the behavior patterns of selected UoIs, might inadvertently capture users who are online simply trying to socialize. These users may not succeed in creating their own network, explaining suspicious feature values such as a low CC and high CIR. Such outliers can end up with a high risk score and rank prominently. This issue could be mitigated by integrating conversation analysis into the system, which is discussed in Chapter 10.2.

Another related issue is the outliers among UoIs who receive a low risk score. Some of these users are well-spoken, and are socially proficient enough to develop a network around them on online platforms. Hence, our system might not be able to efficiently detect these UoIs without the aid of conversation analysis.

Chapter 10
Conclusion

This chapter provides a summary of the key accomplishments of this master's thesis by answering the main research question and subquestions presented in Chapter 1. It also highlights notable insights and suggests potential areas for future research aimed at improving the early detection of cybergrooming.

10.1 Research Questions

10.1.1 Subquestion 1: What features from the user behaviour will best show the difference between normal and predatory/inappropriate behaviour?

The features that best distinguish between normal and grooming/inappropriate behavior include the CC and CIR. The CC helps identify how inter-connected a user's social network is, with lower values often associated with predatory behavior. This is due to them not being a part of a group or organization such as a school class or sports team, where they tend to maintain fewer, less interconnected relationships. The CIR indicates the frequency with which a user initiates conversations. Higher rates potentially indicate users who are overly active in engaging new contacts, which is typical for predatory users.

10.1.2 Subquestion 2: What methods can best be used to evaluate the behaviour of a user such that normal users get scored lower than predatory/inappropriate users?

To evaluate user behavior effectively and ensure that grooming/inappropriate users are scored higher than normal users, we use a FDT. This method involves mapping each user interaction to a specific endpoint within the FDT, based on a set of predefined behavioral features such as the CC, CIR, and the number of neighbors. The FDT is constructed using a series of nested conditional statements that assess these features and guide the interaction to an endpoint that indicates the risk change.

10.1.3 Subquestion 3: What would be a good performance metric for this ranking system, where predatory/inappropriate users are ranked higher than normal users?

A suitable performance metric for this ranking system is the proportion of UoIs among the top-ranked users. For example, in a dataset of 500 users, evaluating the proportion of UoIs within the top 50 users provides a clear measure of the system's effectiveness in prioritizing potentially harmful users.

Additionally, the presence of UoIs at the bottom of the ranking, within a similarly sized subset as the top ranks, serves as another valuable metric. Combining these two metrics offers a comprehensive evaluation of the system's performance. The goal is to ensure that UoIs are not only highly ranked but also that they are minimally present in the lowest ranks.

10.1.4 Main Research Question: Can a user-ranking system based on behavioural analysis streamline the detection of predatory behavior or other inappropriate activities on a platform?

Our work can be seen as a proof of concept, demonstrating the feasibility and potential effectiveness of a user-ranking system for detecting predatory behavior and inappropriate activities online. By analyzing key behavioral features and applying it to a FDT, we identified crucial features such as the CC and CIR, which effectively distinguish between normal and predatory behaviors. The system can accurately rank users according to their risk levels, and performance metrics like the proportion of UoIs among the top-ranked and bottom-ranked users validate its effectiveness.

10.2 Future Research

The work presented in this thesis leaves a lot of potential for future research. One key area for advancement is the optimization of threshold values in the FDT, which could be achieved through the application of machine learning techniques. This process would involve a more extensive effort in labeling the training data to increase its volume and ensure accuracy. Although this would require significant additional work, the result would be an enhanced system with improved precision and reliability. Furthermore, integrating machine learning could help in continuously refining the thresholds based on new data, making the system more adaptive and effective in detecting predatory behaviors.

Another promising approach involves combining conversation analysis with user behavior analysis. By adding branches for chat analysis at specific points in the FDT, the system could be significantly enhanced. This integration would help identify outliers that do not conform to typical predatory behavior patterns identified by our features, such as the user discussed in Chapter 8.4. Merging conversation analysis with behavior analysis would improve the system's ability to detect a broader range of predatory behaviors, making it more adaptive and effective in identifying malicious users early. This approach leverages the strengths of both methods, ensuring that both obvious and subtle indicators of grooming are captured and addressed comprehensively.

References

[15] *Sustainable development goal 16*, https://unric.org/en/sdg-16/#, (Accessed on 06/02/2024), 2015.

[AC16] A. Altay and D. Cinar, «Fuzzy decision trees», *Fuzzy statistical decision-making: theory and applications*, pp. 221–261, 2016.

[AS09] U. Agarwal and U. P. Singh, *Graph theory.* Laxmi Publications, 2009.

[BB19] P. R. Borj and P. Bours, «Predatory conversation detection», in *2019 International Conference on Cyber Security for Emerging Technologies (CSET)*, IEEE, 2019, pp. 1–6.

[BC90] L. Berliner and J. R. Conte, «The process of victimization: The victims' perspective», *Child abuse & neglect*, vol. 14, no. 1, pp. 29–40, 1990.

[BK19] P. Bours and H. Kulsrud, «Detection of cyber grooming in online conversation», in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, 2019, pp. 1–6.

[BSS11] P. Briggs, W. T. Simon, and S. Simonsen, «An exploratory study of internet-initiated sexual offenses and the chat room sex offender: Has the internet enabled a new typology of sex offender?», *Sexual Abuse*, vol. 23, no. 1, pp. 72–91, 2011.

[CC19] R. Chalapathy and S. Chawla, «Deep learning for anomaly detection: A survey», *arXiv preprint arXiv:1901.03407*, 2019.

[EBK95] M. Elliott, K. Browne, and J. Kilcoyne, «Child sexual abuse prevention: What offenders tell us», *Child abuse & neglect*, vol. 19, no. 5, pp. 579–594, 1995.

[Eng23a] I. E. Eng, «Dynamic graph theoretical analysis of cybergrooming detection in chatrooms», M.S. thesis, NTNU, 2023.

[Eng23b] M. Ø. Enger, *Ranking the stars*, Project report in TTM4502, Dec. 2023.

[Fin94] D. Finkelhor, «Current information on the scope and nature of child sexual abuse», *The future of children*, pp. 31–53, 1994.

[GBGZ13] D. Gordon-Messer, J. A. Bauermeister, *et al.*, «Sexting among young adults», *Journal of adolescent health*, vol. 52, no. 3, pp. 301–306, 2013.

[Ham20] W. L. Hamilton, *Graph representation learning.* Morgan & Claypool Publishers, 2020.

[KIF21] A. Kumagai, T. Iwata, and Y. Fujiwara, «Semi-supervised anomaly detection on attributed graphs», in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021, pp. 1–8.

[Lan10] K. V. Lanning, «Child molesters: A behavioral analysis for professionals investigating the sexual exploitation of children», 2010.

[LM12] J. Leskovec and J. Mcauley, «Learning to discover social circles in ego networks», *Advances in neural information processing systems*, vol. 25, 2012.

[LPB09] B. Leclerc, J. Proulx, and E. Beauregard, «Examining the modus operandi of sexual offenders against children and its practical implications», *Aggression and violent behavior*, vol. 14, no. 1, pp. 5–12, 2009.

[McA06] A.-M. McAlinden, «'setting'em up': Personal, familial and institutional grooming in the sexual abuse of children», *Social & Legal Studies*, vol. 15, no. 3, pp. 339–362, 2006.

[Nat24] National Center of Missing & Exploited Children, *Cybertipline® report 2023*, 2024. [Online]. Available: https://www.missingkids.org/content/dam/missingkids/pdfs/2023-CyberTipline-Report.pdf.

[OCo03] R. O'Connell, «A typology of child cybersexploitation and online grooming practices», *Cyberspace Research Unit, University of Central Lancashire*, pp. 6–10, 2003.

[ODER07] L. N. Olson, J. L. Daggs, *et al.*, «Entrapping the innocent: Toward a theory of child sexual predators' luring communication», *Communication Theory*, vol. 17, no. 3, pp. 231–251, 2007.

[RRB22] P. Rezaee Borj, K. Raja, and P. A. Bours, «Online grooming detection: A comprehensive survey of child exploitation in chat logs», 2022.

[Ruh00] B. Ruhnau, «Eigenvector-centrality—a node-centrality?», *Social networks*, vol. 22, no. 4, pp. 357–365, 2000.

[TZHH11] H. Toivonen, F. Zhou, *et al.*, «Compression of weighted graphs», in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 965–973.

[VLA21] M. Vogt, U. Leser, and A. Akbik, «Early detection of sexual predators in chats», in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 4985–4999.

[WEB13] R. Williams, I. A. Elliott, and A. R. Beech, «Identifying sexual grooming themes used by internet sex offenders», *Deviant behavior*, vol. 34, no. 2, pp. 135–152, 2013.

[WJ17] G. M. Winters and E. L. Jeglic, «Stages of sexual grooming: Recognizing potentially predatory behaviors of child molesters», *Deviant behavior*, vol. 38, no. 6, pp. 724–733, 2017.

[WKJ22] G. M. Winters, L. E. Kaylor, and E. L. Jeglic, «Toward a universal definition of child sexual grooming», *Deviant Behavior*, vol. 43, no. 8, pp. 926–938, 2022.

[Zad65] L. A. Zadeh, «Fuzzy sets», *Information and control*, vol. 8, no. 3, pp. 338–353, 1965.

[Aar22] A. F. Aarekol, «A graph theoretical approach to online predator detection», M.S. thesis, NTNU, 2022.

Appendix

Appendix A

A.1 Risk Scores from the Improved System Test

Table A.1: Updated user scores (primary dataset) using the improved system after 1000 interactions.

Rank	User Type	User ID	Risk Score
1	High Risk User	7C8C	2984.0
2	High Risk User	AB7D	2651.0
3	High Risk User	8821	2462.5
4	User of Interest	238F	2200.5
5	High Risk User	A1AB	2036.5
6	User of Interest	CCE0	1763.5
7	Normal User	087F	1646.5
8	High Risk User	399D	1635.5
9	User of Interest	10BA	1297.5
10	Normal User	066E	1259.0
11	User of Interest	207A	1190.25
12	Normal User	08B1	1107.5
13	High Risk User	CBFE	851.3
14	High Risk User	0673	233.8
15	High Risk User	F874	140.5
16	High Risk User	6307	124.5

Table A.1: (continued)

Rank	User Type	User ID	Risk Score
17	Normal User	095E	122.75
18	High Risk User	E3D7	94.5
19	High Risk User	1625	78.0
20	Normal User	0958	68.5
21	Normal User	064E	3.0
22	User of Interest	F83F	0.0
23	User of Interest	647A	0.0
24	High Risk User	1E0B	0.0
25	User of Interest	7B18	0.0
26	Normal User	0636	0.0
27	Normal User	0633	0.0
28	Normal User	0696	0.0
29	Normal User	06DE	0.0
30	Normal User	075D	0.0
31	Normal User	074C	0.0
32	Normal User	073F	0.0
33	Normal User	073E	0.0
34	Normal User	084E	0.0
35	Normal User	0870	0.0
36	Normal User	08F2	0.0
37	Normal User	08BC	0.0

A.2 Risk Changes by Endpoint in Improved System

Table A.2: All risk changes by endpoint in the improved system.

Endpoint	Risk Change
1	-3.00
2	-1.50
3	-1.50

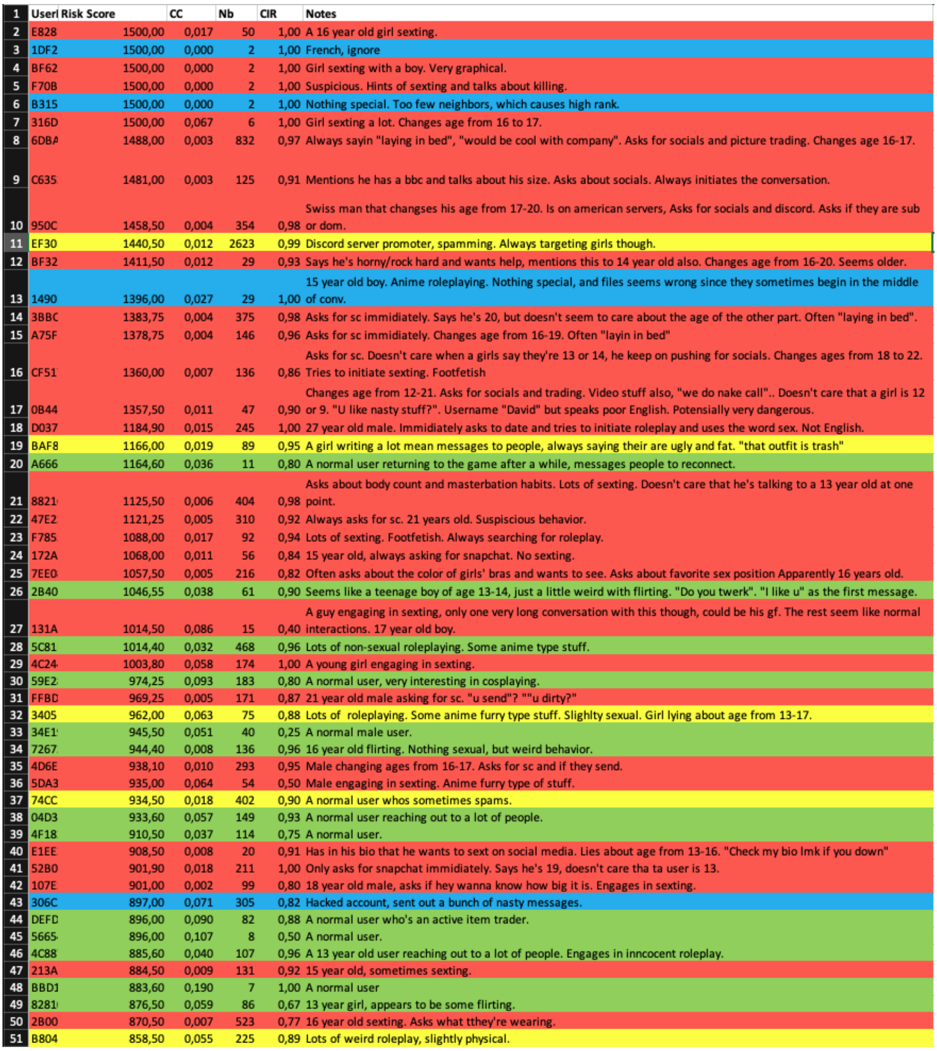
Table A.2: (continued)

Endpoint	**Risk Change**
4	0.25
5	0.00
6	1.00
7	0.80
8	-1.00
9	-1.00
10	1.00
11	1.25
12	1.50
13	1.50
14	0.00
15	0.00
16	1.75
17	2.00
18	2.50
19	3.00
20	3.00
21	-1.75
22	-1.50
23	0.00
24	0.00
25	1.00
26	1.00
27	-0.50
28	-0.50
29	0.50
30	0.75
31	2.00
32	2.50

Table A.2: (continued)

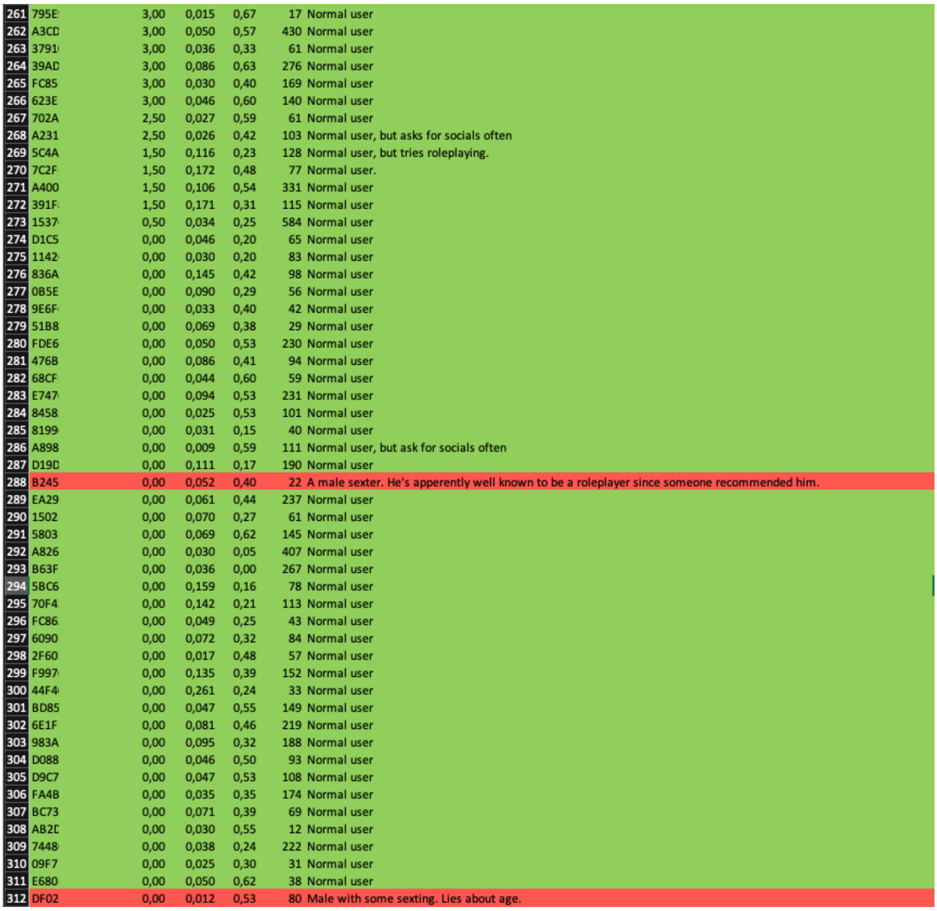
Endpoint	**Risk Change**
33	0.50
34	0.50
35	1.75
36	1.50
37	2.50
38	3.00

A.3 Notes from the Final System Results



1	UserI	Risk Score	CC	Nb	CIR	Notes
2	E828	1500,00	0,017	50	1,00	A 16 year old girl sexting.
3	1DF2	1500,00	0,000	2	1,00	French, ignore
4	BF62	1500,00	0,000	2	1,00	Girl sexting with a boy. Very graphical.
5	F70B	1500,00	0,000	2	1,00	Suspicious. Hints of sexting and talks about killing.
6	B315	1500,00	0,000	2	1,00	Nothing special. Too few neighbors, which causes high rank.
7	316D	1500,00	0,067	6	1,00	Girl sexting a lot. Changes age from 16 to 17.
8	6DBA	1488,00	0,003	832	0,97	Always sayin "laying in bed", "would be cool with company". Asks for socials and picture trading. Changes age 16-17.
9	C635	1481,00	0,003	125	0,91	Mentions he has a bbc and talks about his size. Asks about socials. Always initiates the conversation.
10	950C	1458,50	0,004	354	0,98	Swiss man that changses his age from 17-20. Is on american servers, Asks for socials and discord. Asks if they are sub or dom.
11	EF30	1440,50	0,012	2623	0,99	Discord server promoter, spamming. Always targeting girls though.
12	BF32	1411,50	0,012	29	0,93	Says he's horny/rock hard and wants help, mentions this to 14 year old also. Changes age from 16-20. Seems older.
13	1490	1396,00	0,027	29	1,00	15 year old boy. Anime roleplaying. Nothing special, and files seems wrong since they sometimes begin in the middle of conv.
14	3BBC	1383,75	0,004	375	0,98	Asks for sc immidiately. Says he's 20, but doesn't seem to care about the age of the other part. Often "laying in bed".
15	A75F	1378,75	0,004	146	0,96	Asks for sc immidiately. Changes age from 16-19. Often "layin in bed"
16	CF51	1360,00	0,007	136	0,86	Asks for sc. Doesn't care when a girls say they're 13 or 14, he keep on pushing for socials. Changes ages from 18 to 22. Tries to initiate sexting. Footfetish
17	0B44	1357,50	0,011	47	0,90	Changes age from 12-21. Asks for socials and trading. Video stuff also, "we do nake call".. Doesn't care that a girl is 12 or 9. "U like nasty stuff?". Username "David" but speaks poor English. Potensially very dangerous.
18	D037	1184,90	0,015	245	1,00	27 year old male. Immidiately asks to date and tries to initiate roleplay and uses the word sex. Not English.
19	BAF8	1166,00	0,019	89	0,95	A girl writing a lot mean messages to people, always saying their are ugly and fat. "that outfit is trash"
20	A666	1164,60	0,036	11	0,80	A normal user returning to the game after a while, messages people to reconnect.
21	8821	1125,50	0,006	404	0,98	Asks about body count and masterbation habits. Lots of sexting. Doesn't care that he's talking to a 13 year old at one point.
22	47E2	1121,25	0,005	310	0,92	Always asks for sc. 21 years old. Suspiscious behavior.
23	F785	1088,00	0,017	92	0,94	Lots of sexting. Footfetish. Always searching for roleplay.
24	172A	1068,00	0,011	56	0,84	15 year old, always asking for snapchat. No sexting.
25	7EE0	1057,50	0,005	216	0,82	Often asks about the color of girls' bras and wants to see. Asks about favorite sex position Apparently 16 years old.
26	2B40	1046,55	0,038	61	0,90	Seems like a teenage boy of age 13-14, just a little weird with flirting. "Do you twerk". "I like u" as the first message.
27	131A	1014,50	0,086	15	0,40	A guy engaging in sexting, only one very long conversation with this though, could be his gf. The rest seem like normal interactions. 17 year old boy.
28	5C81	1014,40	0,032	468	0,96	Lots of non-sexual roleplaying. Some anime type stuff.
29	4C24	1003,80	0,058	174	1,00	A young girl engaging in sexting.
30	59E2	974,25	0,093	183	0,80	A normal user, very interesting in cosplaying.
31	FFBD	969,25	0,005	171	0,87	21 year old male asking for sc. "u send"? ""u dirty?"
32	3405	962,00	0,063	75	0,88	Lots of roleplaying. Some anime furry type stuff. Slighlty sexual. Girl lying about age from 13-17.
33	34E1	945,50	0,051	40	0,25	A normal male user.
34	7267	944,40	0,008	136	0,96	16 year old flirting. Nothing sexual, but weird behavior.
35	4D6E	938,10	0,010	293	0,95	Male changing ages from 16-17. Asks for sc and if they send.
36	5DA3	935,00	0,064	54	0,50	Male engaging in sexting. Anime furry type of stuff.
37	74CC	934,50	0,018	402	0,90	A normal user whos sometimes spams.
38	04D3	933,60	0,057	149	0,93	A normal user reaching out to a lot of people.
39	4F18	910,50	0,037	114	0,75	A normal user.
40	E1EE	908,50	0,008	20	0,91	Has in his bio that he wants to sext on social media. Lies about age from 13-16. "Check my bio lmk if you down"
41	52B0	901,90	0,018	211	1,00	Only asks for snapchat immidiately. Says he's 19, doesn't care tha ta user is 13.
42	107E	901,00	0,002	99	0,80	18 year old male, asks if hey wanna know how big it is. Engages in sexting.
43	306C	897,00	0,071	305	0,82	Hacked account, sent out a bunch of nasty messages.
44	DEFD	896,00	0,090	82	0,88	A normal user who's an active item trader.
45	5665	896,00	0,107	8	0,50	A normal user.
46	4C88	885,60	0,040	107	0,96	A 13 year old user reaching out to a lot of people. Engages in inncocent roleplay.
47	213A	884,50	0,009	131	0,92	15 year old, sometimes sexting.
48	BBD1	883,60	0,190	7	1,00	A normal user
49	8281	876,50	0,059	86	0,67	13 year girl, appears to be some flirting.
50	2B00	870,50	0,007	523	0,77	16 year old sexting. Asks what tthey're wearing.
51	B804	858,50	0,055	225	0,89	Lots of weird roleplay, slightly physical.

Figure A.1: Notes and feature values of the 50 highest scoring users in the final system.



261	795E	3,00	0,015	0,67	17	Normal user
262	A3CD	3,00	0,050	0,57	430	Normal user
263	3791	3,00	0,036	0,33	61	Normal user
264	39AD	3,00	0,086	0,63	276	Normal user
265	FC85	3,00	0,030	0,40	169	Normal user
266	623E	3,00	0,046	0,60	140	Normal user
267	702A	2,50	0,027	0,59	61	Normal user
268	A231	2,50	0,026	0,42	103	Normal user, but asks for socials often
269	5C4A	1,50	0,116	0,23	128	Normal user, but tries roleplaying.
270	7C2F	1,50	0,172	0,48	77	Normal user.
271	A400	1,50	0,106	0,54	331	Normal user
272	391F	1,50	0,171	0,31	115	Normal user
273	1537	0,50	0,034	0,25	584	Normal user
274	D1C5	0,00	0,046	0,20	65	Normal user
275	1142	0,00	0,030	0,20	83	Normal user
276	836A	0,00	0,145	0,42	98	Normal user
277	0B5E	0,00	0,090	0,29	56	Normal user
278	9E6F	0,00	0,033	0,40	42	Normal user
279	51B8	0,00	0,069	0,38	29	Normal user
280	FDE6	0,00	0,050	0,53	230	Normal user
281	476B	0,00	0,086	0,41	94	Normal user
282	68CF	0,00	0,044	0,60	59	Normal user
283	E747	0,00	0,094	0,53	231	Normal user
284	8458	0,00	0,025	0,53	101	Normal user
285	8199	0,00	0,031	0,15	40	Normal user
286	A898	0,00	0,009	0,59	111	Normal user, but ask for socials often
287	D19D	0,00	0,111	0,17	190	Normal user
288	B245	0,00	0,052	0,40	22	A male sexter. He's apperently well known to be a roleplayer since someone recommended him.
289	EA29	0,00	0,061	0,44	237	Normal user
290	1502	0,00	0,070	0,27	61	Normal user
291	5803	0,00	0,069	0,62	145	Normal user
292	A826	0,00	0,030	0,05	407	Normal user
293	B63F	0,00	0,036	0,00	267	Normal user
294	5BC6	0,00	0,159	0,16	78	Normal user
295	70F4	0,00	0,142	0,21	113	Normal user
296	FC86	0,00	0,049	0,25	43	Normal user
297	6090	0,00	0,072	0,32	84	Normal user
298	2F60	0,00	0,017	0,48	57	Normal user
299	F997	0,00	0,135	0,39	152	Normal user
300	44F4	0,00	0,261	0,24	33	Normal user
301	BD85	0,00	0,047	0,55	149	Normal user
302	6E1F	0,00	0,081	0,46	219	Normal user
303	983A	0,00	0,095	0,32	188	Normal user
304	D088	0,00	0,046	0,50	93	Normal user
305	D9C7	0,00	0,047	0,53	108	Normal user
306	FA4B	0,00	0,035	0,35	174	Normal user
307	BC73	0,00	0,071	0,39	69	Normal user
308	AB2D	0,00	0,030	0,55	12	Normal user
309	7448	0,00	0,038	0,24	222	Normal user
310	09F7	0,00	0,025	0,30	31	Normal user
311	E680	0,00	0,050	0,62	38	Normal user
312	DF02	0,00	0,012	0,53	80	Male with some sexting. Lies about age.

Figure A.2: Notes and feature values of the 50 of the *lowest* scoring users in the final system.

NTNU

Norwegian University of
Science and Technology