

Identifying Discriminating Age Groups for Online Predator Detection

Submitted for Blind Review

Abstract

Children are a major user group of today's internet, mostly focusing on its social applications like social networking sites. Their online activities are not entirely without risk, as is demonstrated by the presence of online predators, who misuse its potential to search for and contact them with purposes in the area of sexual abuse. In this paper we investigate the presence of these predators on social networking sites and present a method of automatically detecting them. In this process we adopt a genetic algorithm approach that discovers a relation between the amount of friends on an online profile and the age of its owner, searching for statistically significant differences between known offenders and a control group of standard users. After the creation of this model, thresholds are calculated that classify users into danger categories. Experiments on actual police data sources show some promising results in countering this stealthy threat.

Keywords: Online predators, law enforcement, data analysis, discriminating groups

1 Introduction

The growth of the internet and the incorporation of its usage in daily life has been especially strong in the teenage demographic. Children in this age group are young enough to view the internet as an integral part of society, yet old enough to fully utilize its potential, especially in the social area. The onset of web 2.0, the change of the World Wide Web from a static archive of websites into a dynamic environment with interactive programs and websites, has greatly contributed to the way contemporary youth structure their social life. An illustrative example of such a concept would be that of Social Networking Sites (SNS); user profile based dynamic websites that can be used to set-up, maintain and manage networks of (online) friends.

Although the benefits of early accustomization to technology are plenty, internet usage by young people is not entirely without risks, especially since most of the activities are performed alone and usually without adult supervision. Recently there has been a raise in awareness of such dangers and numerous (governmental) campaigns are underway to inform teenagers and their parents of a number of internet hazards, of which the *online predator* is the most prominent. Seen from a general point of view, these predators, "hunt" the internet for other users to exploit, acquiring some asset from them, which can be anything, ranging from monetary gains to identity theft, but most of the time, the term is used to describe the more specific type of *online sexual predator*. These are people that search the internet for sexual prey, mostly children, with the eventual goal of committing some kind of sexual offense. The SNS mentioned

are very well suited for such activities, providing both anonymity and a large, easily searchable and publicly available database of possible victims that are also easily contacted, using the websites standard functionality.

Naturally, this is an area of criminal activity that is currently monitored by police (internet-)detectives. Unfortunately, methods employed by these divisions are mostly passive, dealing with observation of previously encountered online offenders or by reaction to civilian narrative reports on a certain case. It could be a great asset to devise a method that can more actively seek out potential offenders through monitoring systems that detect certain behavioral patterns in an online environment, such as SNS, and classify users of that environment in potential danger categories. This paper aims to provide an initial step toward achieving systems that deal with this issue. It describes the potential ability to detect predators based upon an individual's age and the percentage of under aged "friends" this person has on his (or her) SNS profile. Some of the issues are discussed and a tool for accomplishing this task was tested on actual police data to determine its suitability for and precision in the detection of online predators on SNS.

2 Background

In order to gain a good understanding about the way online sexual predators employ the possibilities of SNS it is imperative to investigate both the (behavioral) properties of these offenders and the main idea behind these interactive websites, providing as much information as needed about the functionality that can be (mis)used within these environments.

2.1 Social Networking Sites

Social networking sites are defined as follows (Boyd & Ellison 2007): social network sites are web-based services that allow individuals to

1. construct a public or semi-public profile within a bounded system,
2. articulate a list of other users with whom they share a connection and
3. view and traverse their list of connections and those made by others within the system.

Thus, being a typical example of Web 2.0 (O'Reilly 2005), users are the main source of content within SNS, being responsible for its creation and maintenance, where the service provider is only responsible for the development and maintenance of the framework. In practice this means that people can register at the SNS by using a nickname (username, profile-name) after which they can build a personal page, sometimes with a picture of him or herself. Physically, this page is located on a server of the SNS and

can be setup and changed by the user through a *content management system*, which eliminates the need for knowledge of web languages like HTML. Usually, information like hobbies, geographical location, work and personal situation is added to the page which enables other SNS users or, in the case of a public profile, all internet users to check if the newly registered individual has common interests or lives in the same country or area. It is common for a SNS to provide the possibility to limit the access to a certain profile to SNS members or even to a group of specified people

As stated above, the core of a SNS is the possibility to create links between one profile and a (large) number of other profiles to state that a certain connection exists between the owners of both profiles. These connections can vary from real-life friends, online friends to club memberships and corporate employments. The nature and nomenclature of these connections may vary from site to site, but in this paper we will refer to these connections as *friends* of the profile holder. Note, that it is often not possible to link to profiles on different SNS, effectively limiting users to have friends within the same framework alone. The list of friends is usually publicly viewable so profile visitors can immediately see who the (other) friends of that specific person are. It is not always the goal of a SNS user to meet new people. A lot of SNS users only use the SNS to communicate with people they already know and also have contact with in real life. There are also SNS users who add almost anybody to their friends list with a result of having hundreds or even thousands of friends.

Most SNS provide a search option that allows users to search for people they know, for example a search by name, or for profile holders that adhere to a specific set of properties specified by the searching party. Sometimes, the friend list of other users can be searched in this manner as well. Within the scope of this paper it is noteworthy that a search on age option is common, providing an easy way to specify certain “preferences” a predator might have when searching for potential victims.

2.2 Online Identities

One of the problems in today’s internet that certainly holds within SNS is that of “identity”. In real life, identity is defined as (American Psychological Association 2007):

“The distinguishing character or personality of an individual.”

This definition cannot be translated directly to an online environment, where it is possible to maintain a profile that resembles your offline identity most closely or to craft an entirely new profile that suits a specific purpose, for example when applying for jobs, meeting people of the other gender or, in the case of the online predator, to mask one’s true identity to reach out more easily to potential victims. When communicating in the physical world the human body is automatically connected to identity. When talking to one another, people can therefore easily confirm that the person who they are communicating with, is really who he says he is (Donath 1998). This is one of the reasons why communication on the internet and especially on SNS, where physical recognition is limited to visual media chosen by the user under observation himself, can be easily used to mislead people. Also, the inability of detectives to successfully match all known offenders to online profiles poses a significant surveillance issue. Therefore, the difference between the physical identity and the

online identity, plays a crucial role in criminal cases related to SNS.

The difference between offline and online identities can have advantages as well. Law enforcement agencies around the world can use covert identities, feigning for example that they are a 12 year old boy that wants to meet the individual under observation in a hotel. On arrival the offender can then be arrested by law enforcement officials. Especially in the United States methods like these have been successfully employed (Mitchell et al. 2005), but they are often outlawed in other countries due to entrapment issues. In these countries a more passive monitoring approach with early warning systems in place would probably yield the best results in countering the threat of online predators. A good starting point would be to analyze the demographics of these offenders.

2.3 Online Sexual Predator

Comparable to the animal world, a predator in the human world “hunts” other individual for some commodity, in the process hurting the other person’s interests, identity or even life. The scope of this paper is limited to the area of sexual predation, where the modus operandi is set in an online environment. In literature, both the term online sexual predator and online predator is used to describe the same phenomenon. Although a small percentage of victims concerns adults, the online predator mostly preys on minors or under aged individuals, which is the type of predation this paper focuses on.

Child sexual predators are a heterogeneous group, and as a result it is difficult to define a typology of the sexual predator (Dombrowski et al. 2007, 2004). Psychologists and law enforcement investigators employ the following, somewhat crude definition (Elliott et al. 1995), which claims a typical sexual predator is likely to be:

- a male,
- aged 18-72 (preponderance between 30 and 42),
- having a successful career (middle to upper management),
- being a college graduate,
- being married or divorced,
- being a parent himself (but whose children are older than his targets).

It is plausible that online sexual predators are younger of age compared to the traditional sexual predators mainly because of the usage of modern technology. Research done in this area tends to support this (Sullivan 2002, Finkelhor et al. 2000). When looking at the above mentioned figures, we can assume the typical online sexual predator is a male, aged between 18 and 65.

In contrast to the standard “*grooming*” approach, the process that bridges the gap between being strangers and having intimate physical contact, online sexual predators mostly use non-traditional ways to get in contact with children, searching the internet on chat rooms and websites specifically meant for minors. While searching, the predator tries to find individuals who fit the age, sex and location of their preference. When a potential victim is found, contact is usually made through chat (Malesky 2007). A likely next step in the online grooming process is a potential inclusion in the predator’s network on a SNS.

3 Predator Presence on Social Networking Sites

To analyze activities of online predators on SNS it is best to focus at a specific SNS. A lot of the SNS basically work similar, providing the same structure and functionality, therefore results yielded by the following research are assumed to be valid for most SNS. In this paper we chose to investigate the presence of online predators on the largest and most used SNS in the Netherlands: Hyves. This SNS is used by people of all ages with a strong group of people under 35.

To get an impression of the amount of online predators being active at Hyves a list of known offenders was extracted from the national database of the Child Pornography Team of the Dutch National Police Agency. To make the list manageable, only offenders were added who are known for being active on the internet and fall within the demographics described above; consequently the offenders in this group can be called potential online predators. This shortlist served as a basis upon which a brute force search was performed by hand to match these individuals with a high degree of certainty to a Hyves-profile, comparing the data on first name, last name, home town, age and date of birth. All possible matches with a slight degree of doubt were omitted from the result list. Obviously, people with a fictional profile name, address, age, etc were either not identifiable or excluded due to doubt, so the quantifications displayed in Figure 1 can be viewed as a lower bound of predator activity levels on this SNS.

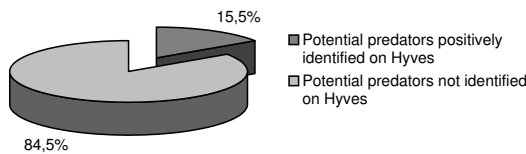


Figure 1: Predator Activity on Hyves based upon shortlist

Figure 1 suggests that at least 15% (316) of 2044 child sexual predators on the shortlist is active on this specific SNS alone. This situation poses a significant risk for children using the same site as well. This is confirmed by the increasing number of narrative reports and reports filed at different hotlines (See Figure 2).

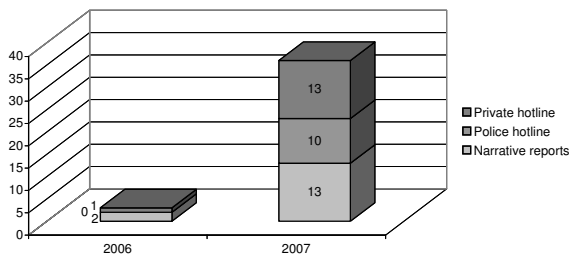


Figure 2: Increase of reports on predator activity on Hyves

A number of reports indicated that a parent ran into a profile of an unknown adult on his or her child's *friend list*, noticing that this person's profile contained a lot of other minors as well. Although, logically, no legal action can be taken against such individuals, this phenomenon gives rise to the assumption that a certain percentage of under aged friends can be a good indication of an individual's chance of falling in a certain danger category.

3.1 Amount of under-aged friends: A first glance

As a first step in the automatic discovery of such a relationship, the total amount of friends and the number of minors on that list was retrieved for every offender on the shortlist. In addition, 960 random Hyves users were selected (10 individuals, of both genders, of every age between 18-65), who serve as a control group in this experiment. The same information was retrieved for this group, which resulted in Figure 3.

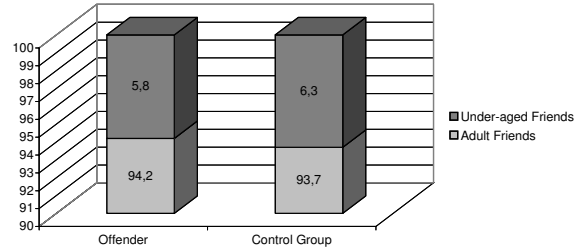


Figure 3: Average percentage of under-aged friends

At a first glance, the data represented in this figure rejects the current hypothesis, showing no significant difference and even a slightly larger percentage of under-aged friends for the control group. However, a more thorough investigation of the offender group reveals the reason for this perceived equality in variance: offenders in a lower aged sub-group (18-25) are underpresent compared to their control group counterparts, while the percentage in this group is relatively high compared to higher age groups. Hence, a more detailed analysis is warranted.

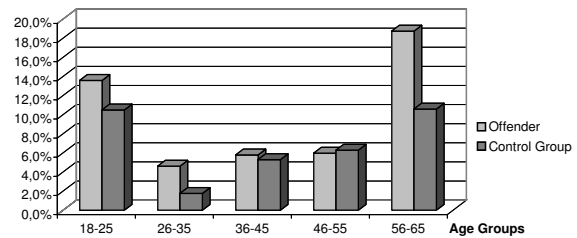


Figure 4: Average percentage of under-aged friends per age group

Figure 4 clearly shows that in the majority of the age groups the offender group has a larger percentage of under aged friends. However, not all the differences can be designated as significant. Furthermore, the arbitrary nature of the currently investigated age groups, makes it hard to draw any conclusions about the possibility of danger category classification based upon this property of a SNS profile. A method of automatically generating and evaluating age groups with a significant difference would greatly contribute to this cause. Moreover, the aggregation of threshold values for classification out of these discriminating groups has the potential to greatly improve police monitoring activities.

4 Approach

The percentage of under-aged friends varies greatly between the different age groups, both in the offender and the control group as is demonstrated in Figure 5. However, the graph clearly reveals a large schism between the percentages of the two different groups for some ages. An age-wise comparison is therefore necessary to examine a possible relation between the two.

During this process, age groups can be constructed that clearly discriminate between the two groups, due to the significance of the difference in percentage in their sample populations.

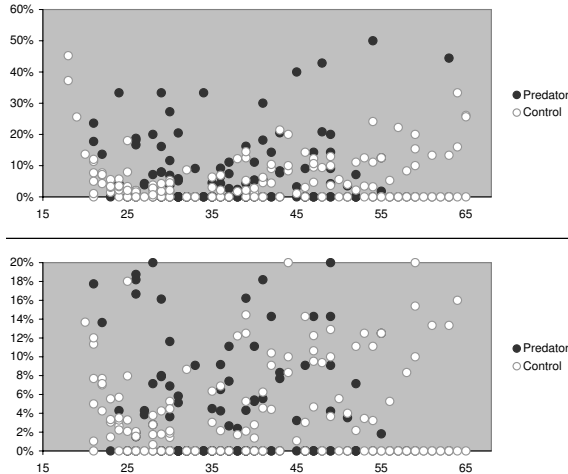


Figure 5: Scarce scatter plot for separate ages

There are two properties of the experiment that make the recognition of such groups non-trivial. First, not one of the ages supports samples large enough to validate a statistical test on significance for this single age. Hence, no conclusion can be drawn about the discriminative properties for a single age and grouping must occur on this variable before significance can be established with any certainty.

Second, the variance within groups for the age variable is large, which makes the grouping of single ages for the establishment of a discriminative property difficult. Furthermore, it is not always clear beforehand what effect this variance will have on the discriminative property of two groups (possibly of size 1) that are about to be merged for evaluation. An example of this can be seen in Figure 6, where both ages show a clear separation between the control and offender group, but merging might or might not lead to a group that retains this significance in difference because of “overlapping”. A special search method for these combined age groups is therefore warranted.

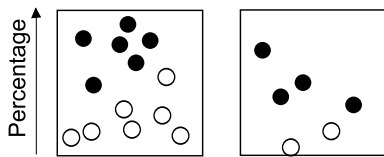


Figure 6: Percentages for both groups might overlap if merged

For this purpose we propose a simple *genetic algorithm* that attempts to identify age groups that are both as large as possible and satisfy the significant difference requirement. After a certain amount of evolution cycles, the process yields a number of disjunct subgroups that satisfy the significance demand. They should contain the age groups that offer the best discrimination between the offender and control group and can be used for classification of new individuals. At the end of our approach (see Figure 7), the results will be tested against itself through a ten-fold cross validation method.

4.1 Genetic Algorithm

A genetic algorithm is a programming technique that mimics biological evolution as a problem solv-

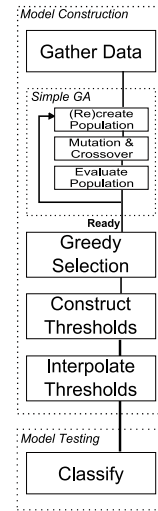


Figure 7: Approach

ing strategy. Its foundations lie in a number of potential solutions (*candidates*) to that problem, and a metric called a *fitness function* that evaluates each candidate quantitatively, resulting in its *fitness*. Initial candidates are often generated randomly. After evaluation, all promising candidates are kept and allowed to reproduce. This is done based upon some pre-established operators, that most often include *crossovers*, combinations of existing candidates, and *mutations*, which are copies of existing candidates with slight modifications. Again, the most promising candidates are kept and the cycle continues in this manner for a large number of generations. The expectation is that the average fitness of the population will increase each generation, eventually resulting in excellent solutions to the problem under observation.

After the data selection, which was discussed in Section 3.1, candidate age groups were selected as starting population of the genetic algorithm. All possible combinations of two consecutive ages were considered candidates, resulting in a population of 46 candidates. An *elitist* selection method was adopted in order to preserve any older candidates with higher fitness than that of the newly generated candidates. Although this increases the risk of obtaining a *local optimum*, a situation where small steps backwards, needed for a potential large step forward, are immediately discarded, it speeds up the algorithm by preserving good, already realized, candidates that could potentially be lost by a series of unfortunate operator applications.

4.1.1 Operators

In every generation, the fittest candidates will be used to construct the new generation in three distinct ways, two of which are mutations and one that is a crossover.

The first mutation involves exchanging one particular age in the selected candidate group with a randomly selected other age, that is currently not in the group. Naturally, this will not affect the size of the group under observation. During this process, a copy of each of the before mentioned candidates is subjugated to this mutation resulting in 46 extra candidates.

The second mutation either removes a random age from a copied group or adds a random age to this group. Both options have an equal chance of being selected. Note that this mutation changes the size of the group, increasing or decreasing it by 1. This

mutation also introduces 46 new candidates to the population. Note that no group is ever reduced to size 1 or lower.

The crossover operator randomly selects couples of candidates that will be merged into one new candidate. Each candidate will be used for this crossover, resulting in 23 new candidates, and the process is performed twice so that the size of the total population after crossover is exactly 4 times that of the original (46 \rightarrow 184). Note that candidates matched for merging are not necessarily disjunct. Since an age can not be present in an age group twice, the size of the merged candidate is not always the sum of the sizes of its “parents”.

Together, the mentioned operators provide a wide range of possibilities for improvement, providing fast growth of successful age groups through crossover, slight increases in size for good solutions and slight mutations within a fixed group for solutions that approach perfect.

4.1.2 Fitness Function

The selection method chosen to judge the fitness of all individuals in the population is two-fold. First there is a hard demand, stating that all age groups must show a significant difference between the control and offender group. Second, the larger the groups are, the more accurately, and according to Occams Razor, the most plausibly they describe the online predator.

This selection is realized by assigning a bonus $\mathcal{B} = 10$ to an age group if its significance property has reached a certain threshold. This way, candidates that satisfy the significance demand are almost always ranked fitter than groups that do not.

Significant Difference

A significance test is a formal procedure for comparing observed data with a hypothesis, whose truth is to be assessed. The results of such a test are expressed in terms of a probability that measures how well the data and the hypothesis agree. Usually, the difference between two populations on some variable is calculated by assuming such a difference does not exist, called the *null-hypothesis* or \mathcal{H}_0 , followed by a calculation that significantly falsifies that assumption, investigating the strength of the evidence against the hypothesis in terms of probability. The computed probability that denotes the chance that \mathcal{H}_0 is true is called the *p-value*. The smaller, the *p-value*, the stronger the evidence against \mathcal{H}_0 is. (Moore & McCabe 2003). A number of different techniques exist that perform the calculation of this value, the *Student’s t-test*, or “*t-test*” for short, being the most common. If the *p-value* is smaller than a certain significance level, denoted by α , which is usually set at 0.05, the groups are said to vary significantly on this variable.

There are a number of different *t-tests* of which the *t-test* for the comparison of two independent samples is applicable in our approach. In this test, the means (\bar{x}_c and \bar{x}_o) and standard deviations (s_c and s_o) of the sample populations are tested against the respective theoretical means (μ_c and μ_o) and standard deviations (σ_c and σ_o) of their entire population, where *c* and *o* denote the control and offender group respectively. Through inference on $\mathcal{H}_0 : \mu_c = \mu_o$ a *t-value* is calculated as follows:

$$t = \frac{\bar{x}_c - \bar{x}_o}{\sqrt{\frac{s_c^2}{n_c} + \frac{s_o^2}{n_o}}},$$

where n_c and n_o are the respective sample sizes. This

t-value effectively denotes the amount of theoretical standard deviations \bar{x}_c lies from \bar{x}_o . This *t* has a *T-distribution*, a bell-shaped, 0-centered distribution that approaches the normal distribution when its “*degrees of freedom*” (*df*), approach infinity. This number is usually set to $n_c + n_o - 2$. The *T-distribution* lies just above the normal distribution in the outer corners, hence the more degrees of freedom, the lower the distribution will lie, the greater the chance that the *p-value* be lower than α , significantly rejecting \mathcal{H}_0 . The calculation of the *p-value* is shown in Figure 8.

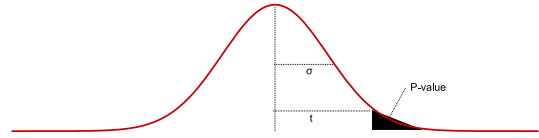


Figure 8: Calculating the P-value based on *t*

The *p-value* can be calculated by integrating on the *t-distribution*, with parameter *df*. The specification of this function falls outside the scope of this paper but can be used as follows where \mathcal{T} is the *p-value* reached through the *t-test* and *tdist* is the integrated function for the *t-distribution* with *df* degrees of freedom:

$$\mathcal{T} = \text{tdist}(t, df)$$

A difficulty that arises with the usage of the *t-test* is that it assumes *normality*, meaning that the populations are normally distributed on the variable under investigation. This normality is said to hold in practice if the variable is distributed normally or if both sample population sizes are at least 30. Within our approach, the first demand is certainly not satisfied, while the second demand only holds for some of the larger age groups. Consequently, we have to resort to an alternative when either n_c or $n_o < 30$.

One of these good alternatives is the *Mann-Whitney test* (MW). This is a robust method that strives to compute a *p-value* for all populations, even those with a small size. Naturally, some accuracy is sacrificed reaching this objective. MW is based on quantitatively ranking every person in the sample based upon the variable under observation, followed by calculating the sum of ranks (\mathcal{S}_x) and average rank ($\bar{\mathcal{S}}_x = \frac{\mathcal{S}_x}{n_x}$) per group ($x = c$ or $x = o$). If $\mathcal{H}_0 = \mu_c = \mu_o$ holds, it is to be expected that $\bar{\mathcal{S}}_c \approx \bar{\mathcal{S}}_o$.

The test involves the calculation of a statistic, usually called \mathcal{U} , whose distribution under \mathcal{H}_0 can be approximated by the normal distribution. This is done by comparing the realized sum of ranks for one group (\mathcal{R}_x) with its maximal value:

$$\mathcal{U}_x = n_c n_o + \frac{n_x(n_x - 1)}{2} - \mathcal{R}_x,$$

where *x* is either *c* or *o*. Then

$$\mathcal{U} = \min(\mathcal{U}_c, \mathcal{U}_o)$$

Now *z*, the mount of standard deviations both group’s mean differ, comparable to *t* from the *t-test*, can be computed by

$$z = \frac{(\mathcal{U} - \mu_U)}{\sigma_U}$$

μ and σ being the mean and standard deviations respectively, where, supposing \mathcal{H}_0 is true

$$\mu_U = \frac{n_c n_o}{2}$$

and

$$\sigma_U = \sqrt{\frac{n_c n_o (n_c + n_o + 1)}{12}}$$

The p -value can be calculated by integrating on the normal distribution. Where \mathcal{Z} is the p -value reached through the Mann-Whitney test and $ndist$ is the integrated function for the normal distribution:

$$\mathcal{Z} = ndist(z)$$

Combining both methods in such a way that the strengths of both are used on applicability, statistical significance is computed as follows:

$$\mathcal{P} = \begin{cases} \mathcal{T} & \text{if } n_c \geq 30 \text{ and } n_o \geq 30 \\ \mathcal{Z} & \text{otherwise} \end{cases}$$

Combined with size

Now that the significance of the difference in amount of under-aged friends is computed, the group size variable can be added to calculate the fitness of an age group. A simple, yet effective, way to denote the size of a group is to count the number of single ages present (n_g). Measuring group size in individual ages also matches well with the computation of significance, which will be normalized between 0 and 10. Now, the fitness \mathcal{F} of an age group is represented by:

$$\mathcal{S} = \begin{cases} \mathcal{B} + 10 \cdot \frac{\alpha - \mathcal{P}}{\alpha} & \text{if } \mathcal{P} \leq \alpha \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{F} = n_g + \mathcal{S},$$

where \mathcal{S} is the computed significance. This function yields a fitness between 2, being the minimum group size, and $\max(\mathcal{N}_g) + 20$.

After applying the fitness function to all candidates, the population is ranked accordingly. Only the best quarter of the population is kept as the new generation, resulting in again 46 candidates.

The genetic algorithm presented above can end (successfully) in two ways. Either the average fitness has not increased for a set number of generations, most often signifying that no ‘‘child’’-candidates are created that perform better than their ‘‘parents’’, or a maximum number of generations is reached without entering such a state, after which the best candidates are selected anyway. This last number is usually reasonably large, ranging from 100,000 to 1,000,000 depending on the problem at hand.

4.2 Greedy Group Selection

After the GA finished, there are 46 groups remaining that have the potential of realizing the goal of discriminating between the two groups based upon age and percentage of under aged friends. We propose a greedy selection method that extracts the best candidates for this task from the final generation. There are six steps in this process

1. Sort candidates on fitness,
2. Start at the best candidate,
3. Scrap the candidate if it does not show a significant difference,
4. Select the candidate if it is disjunct to all previously selected age groups,
5. If not disjunct, cut the intersection and re-evaluate and resort all items ranked below including this item,
6. Go to next, lower ranked candidate and continue at step 3.

The first step is a logical step to take since significance is the goal to be reached, but after a large number of evolution cycles this step will be used very infrequently.

As is illustrated in Figure 9, an overlap can occur between two different age groups in the final selection. As step 5 dictates, the group with the lower fitness score surrenders the intersection. There are two reasons for this. First, as was demonstrated in Section 4 there can be a different boundary or threshold (Section 4.3) between control and offender group which prohibits a merger of the two if such a situation arises. If such a merge would have been possible, chances are that it would have happened during the GA phase of the approach. Second, the age group selected first is preserved, because for some reason it had a higher fitness than the second group. This can either be because it shows a more significant difference, is larger or both. In all these situations preservation of the stronger candidate is the preferred option.

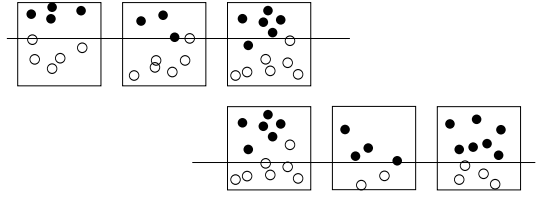


Figure 9: Two overlapping age groups, with different thresholds

Naturally, the size of the second group will decrease, reducing its fitness. One might also consider the chance that its significance could potentially be lowered, however, chances of this happening are extremely slim. Due to the fact that both groups were not merged in the GA phase, it is safe to assume that their boundaries differ. Hence, removal of the overlapping age will only set the threshold even further away from the first group, more closely approximating the boundary for the remaining ages and removing potentially overlapping percentages between the ages in the group. This will only lead to a higher significance level. Eventually, the fitness of the second individual does not necessarily fall greatly due to this modification.

4.3 Thresholding

Now that we have selected the discriminating age groups with the most descriptive value, they can be employed for classifying individuals in danger categories. Before this can be accomplished a threshold \mathcal{C}_g must be designated for each age group g that denotes the decision line above which an individual is classified as potential offender and under which can be seen as safe. There are a number of different ways of constructing such thresholds, from which we choose the following four:

- Least false positives,
- Least false negatives,
- The average of the above.
- The weighted average of the above.

If the first option is applied, the method operates on the safe side, preferring to miss a predator, rather than falsely including a regular user. The second option can be used if one hopes to detect all predators, falsely classifying some regular users in the danger group. Taking the average of both could provide a more balanced view of reality, allowing for some false positives and negatives. The weighted average shifts the line more toward the least false negatives line if

there were relatively many offenders in this age group and vice versa. The rationale behind this is that the more individuals were present in a sample the more representative that sample is for the entire population. It is calculated as follows:

$$\text{weightedaverage} = \frac{\frac{n_o}{n_c} b_{\text{lfm}} + \frac{n_c}{n_o} b_{\text{lfp}}}{2},$$

where b_{lfm} and b_{lfp} are the thresholds calculated by option 2 and 1 respectively.

It can also be the case that both groups are separated strictly, which occurs quite often if our hypothesis is correct. If this holds for a certain age(group), the method of setting the threshold is less relevant, but an average of the first two options is still preferred for clarity reasons.

Also, there are two options of setting these thresholds, by group or by every age from the groups separately. The first has the advantage that a clear image is communicated about the nature of the predator, while the second is more accurate, reducing the need for approximating over a larger number of samples.

It might be interesting to investigate how the area between b_{lfm} and b_{lfp} , that we will call the *ribbon*, is populated. Calculating the division of offenders and control and the total amount of items in the ribbon can indicate how many overlap there is between the two samples. The ribbon \mathcal{R} is calculated by:

$$\mathcal{R} = b_{\text{lfm}} - b_{\text{lfp}}$$

If there is any overlap between the sample populations the ribbon size will be negative, since the “no false positive”-line will lie under some of the samples in the control group. Therefore, a positive size of the ribbon is desirable, its size depicting the confidence of the existing significance.

The thresholds chosen within this phase can potentially be used to interpolate (for example through linear regression), to set a threshold for all ages, even though they were not shown to have a significant difference. This could both validate the exclusion of these ages, if a low accuracy is reached during testing and provide a function that can be used in future applications reducing storage needs of the outcome of this research.

5 Experimental Results

As mentioned in Section 2 our approach was tested on the national database of the Child Pornography Team of the Dutch National Police (KLPD) and a randomly drawn sample population of Hyves-users. We used all standard settings for our variables setting $\mathcal{B} = 10$, $\alpha = 0.05$, $df = n_c + n_o - 2$ and set the maximum number of evolution cycles to 100,000.

Figure 10 displays all the age groups that were discovered with the proposed approach. They are drawn at the weighted average threshold over entire age groups. In total, 4 different age groups were discovered, of which 3 were consecutive and 1 consisted of two separate groups. Together they contain 60.4% of all ages. Naturally, this graph has approximately the same form as Figure 4.

During the ten-fold cross-validation we computed the accuracy of our approach with all the different thresholding settings specified above. It was calculated by dividing the number of errors made in classification (both false positive and false negatives) by the number of correct classifications and subtracting that from 100%. Table 1 shows the results for all thresholding settings:

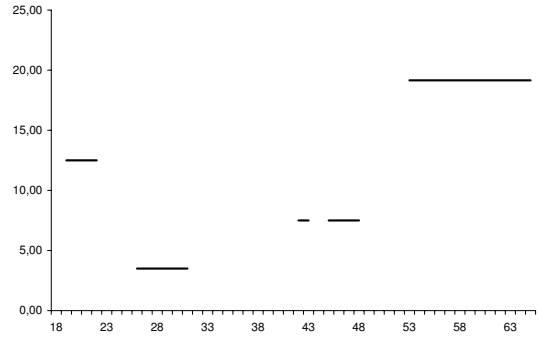


Figure 10: Discriminating age groups drawn at threshold height

Table 1: Accuracy results for all types of group thresholding

	<i>LFP</i>	<i>LFN</i>	<i>Average</i>	<i>W. average</i>
Accuracy	78%	81%	89%	92%

As discussed in Section 4.3, classification thresholds can also be set for individual ages instead of entire discriminating age groups. This could potentially lead to a higher accuracy. It also opens the door to interpolation of the results through linear regression. Figure 11 shows the results from this endeavor.

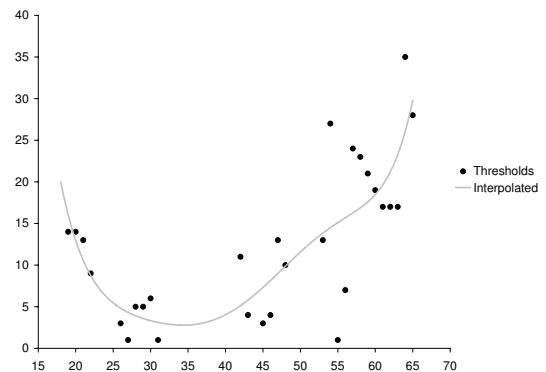


Figure 11: Individual thresholds supporting an interpolation line

It is clear from the graph that there is still a large variance in thresholds between the different “members” of an age group. This leads to the assumption that even better accuracy can be reached if the individual members are treated separately. If thresholding is done on an individual basis the results of an accuracy test differ quite a bit from above, as is obvious from table 2. Also, now that the thresholds are points in the graph, an interpolation could be performed that set thresholds based upon a polynomial, which yielded results that are also presented in the table.

Table 2: Accuracy results for all types of individual thresholding

	Discriminating groups					Other
	<i>LFP</i>	<i>LFN</i>	<i>Avg.</i>	<i>Wei.</i>	<i>Int.</i>	<i>Int.</i>
Acc.	83%	84%	95%	98%	82%	52%

The small size of the ribbon can be clearly ob-

served in both the group thresholding and the individual thresholding case. It is 0.28 and 0.39 percentage points on average respectively. The items in the ribbon are 69% individuals from the offender sample.

6 Conclusion and Future Directions

The test shows a surprisingly large percentage of ages that actually support a significant difference between the offender and control samples. For these groups the best way to set a classification threshold appears to be based upon individual ages within this group. A weighted average of the “least false positives” and “least false negatives” extremes yields an accuracy of 98%, which suggests the method is a reliable tool to predict if a certain profile holder on a SNS is an online predator.

Unfortunately, at the same time, it appears to be impossible to draw any conclusion about people falling out of the discriminating age groups, only if this is only 40% of the population. An accuracy of 52% is too low to classify anything with any certainty. It is unfortunate that the preponderance interval of predators (30–42, Section 2), falls in this undecidable interval for 70%. However, the fact that no classification can be done after interpolation suggests these ages were left out by our method correctly, having no discriminative features. Interpolation is also to be avoided because it adversely affects the accuracy within the discriminating age groups.

The fact that the ribbon is very small suggests that the significance of the difference reached within the age groups barely holds. On a positive side, the number of offenders in the ribbon is substantially larger than the number of individuals from the control group. This indicates that our method classifies more false negatives than false positives, which is a good thing in most judiciary systems, supporting innocence until proven otherwise.

Now that these discriminating age groups have been established, they can be efficiently used within monitoring or early warning systems. Only 23 combinations of age and percentage need to be stored in order to do the classification.

Future research should mainly focus on two things: first, the outcome of our approach should be tested on other SNS, which can be easily done depending on the availability of police data, and second, the approach could be tested for other variables than age and percentage of under aged friends. The latter can be used to validate the outcome of our approach and could potentially shed more light on the age groups that were undecidable in our tests.

A last difficulty we ran into is the fact that not all queries can be executed via the openly accessible Hyves website. This means that for a detailed search query access to the full Hyves database is needed which is not possible for the ordinary Hyves user. After consultation, Hyves decided not to approve usage of its database directly, other than by official police demand, which we did not have. This severely limited a final test, that could have shown numbers of detected potential predators that were not on our shortlist or revealed profiles that we encountered in the construction of that list that were rejected because of doubts. If privacy legislation allows it (See Section 6.1), future work could aim at acquiring this permission or extracting this information from the Hyves site by using web spiders in order to perform a final test that can provide a complete test case and an actual risk analysis on the presence of predators on SNS.

6.1 Privacy Concerns

Naturally, with the classification of unknown individuals based upon public profiles of potentially alternate identities for the purpose of law enforcement come some huge privacy and judicial concerns. Although the results seem promising, the volatile nature of statistics prohibits usage of our methods in proceedings of law. It may however serve its purpose as a monitoring tool for larger internet applications, identifying profiles that may pose a risk to children and could be reviewed by a detective, but the possibilities in this area heavily depend on judicial constraints within the country of origin or application. In contrast to the more active approaches described above, the method has also potential as a strategic tool, helping to chart the dangers of predators on SNS that could influence future legislation or police priorities. This usage does not necessarily translate general acquired knowledge to individuals and is therefore less prone to privacy concerns.

During our research no “new” potential predators were identified nor put in any list that is currently monitored by police officials. Also, all data used within the research was either publicly available on the internet or made available by the police in an anonymized version. Hence, no privacy sensitive footprints of our research remain.

References

- American Psychological Association (2007), *Merriam-Webster's Medical Dictionary*, Merriam-Webster, Incorporated.
- Boyd, D. & Ellison, N. (2007), ‘Social network sites: Definition, history and scholarship’, *Journal of Computer-Mediated Communication* **13**(1).
- Dombrowski, S., Gischklar, K. & Durst, T. (2007), ‘Safeguarding young people from cyber pornography and cyber sexual predation: a major dilemma of the internet’, *Child Abuse Review* **16**(3), 153–170.
- Dombrowski, S., LeMasney, J., Ahia, C. E. & Dickson, S. (2004), ‘Protecting children from online sexual predators: Technological, psychoeducational, and legal considerations’, *Professional Psychology: Research and Practice* **35**(1), 65–73.
- Donath, J. (1998), ‘Identity and deception in the virtual community’, *Communities in Cyberspace*.
- Elliott, M., Browne, K. & Kilcoyne, J. (1995), ‘What offenders tell us’, *Child Abuse and Neglect* **19**, 579–594.
- Finkelhor, D., Mitchell, K. & Wolak, J. (2000), Online victimization: A report on the nation's youth, Technical report, National Center for Missing and Exploited Children.
- Malesky, L. (2007), ‘Predatory online behavior: Modus operandi of convicted sex offenders in identifying potential victims and contacting minors over the internet’, *Journal of Child Sexual Abuse* **16**(2).
- Mitchell, K., Wolak, J. & Finkelhor, D. (2005), ‘Police posing as juveniles online to catch sex offenders: is it working?’, *Sexual Abuse: A journal of research and treatment* **17**(3).
- Moore, D. & McCabe, G. (2003), *Introduction to the Practice of Statistics*, 4 edn, W.H. Freeman and co.

O'Reilly, T. (2005), 'What is web 2.0 - design patterns and business models for the next generation of software'.

Sullivan, M. (2002), *Safety Monitor, How to Protect your Kids Online*, Bonus Books.