SafeChat: Using Open Source Software to Protect Minors from Internet Predation

Brett Thom

Departments of Mathematics and Computer Science

Mentor: Dr. April Kontostathis

Spring 2011

Submitted to the faculty of Ursinus College in fulfillment of the requirements for Distinguished Honors in Computer Science

Distinguished Honors Signature Page

Brett Thom

SafeChat: Using Open Source Software to Protect Minors from Internet Predation

Advisor: _____

April Kontostathis

Committee Members:

_____

April Kontostathis

_____

Akshaye Dhawan

`

_____

Lynne Edwards

Outside Evaluator:

_____

John P. Dougherty

Approved: _____

Roger Coleman

Abstract

This thesis describes the development of a plugin for Pidgin, an open source instant messaging client. The plugin, SafeChat, allows parents to protect their children from Internet predators. SafeChat gives parents a better alternative than other currently available tools by providing a communication-theory based tool.

To achieve its task, SafeChat first extracts the chat data from the Instant Messages (IMs). After the data is successfully extracted and put into a file system, detection algorithms are run on the data to classify the chat participants' behaviors as predatory or non-predatory.

Contents

Chapter 1: Introduction

Sexual predation is defined as an attempt by an adult to engage a minor in sexual activity. The Internet provides a means for perpetrators to meet potential victims, and the number of children targeted is continuing to grow. According to a study by Wolak, Mitchell, and Finkelhor that occurred over a five year period, unwanted exposure to sexual material increased 9%. Forty percent of youths said they were asked to send a person nude or sexual pictures of themselves. Thirteen percent of the youth said that they were solicited online and 31% of those solicitations were for offline contact [1].

Only a minority of these minors said that these incidents were distressing [1]. It should be asked though; can a 12 year old really be expected to discern the degree between right and wrong? Responsible adults and parents need to do what they can to protect minors. SafeChat is a tool that parents could use for this purpose.

Chapter 2: Background

2.1 The Environment

Online Instant Messaging is a system that uses synchronous text chat involving point to point communication between two users. Instant Messaging (IM) systems also allow users to have group chats; these can either be entirely public or can be restricted to a controlled group of users. IM allows people to converse with complete strangers from around the world [2].

Teenagers and other minors are relying on and using Instant Messaging services more and more. Part of the reason for the widespread growth of this technology is peer pressure. Many minors want to use IM services because their friends use IM. Minors believe they will be socially out-cast from a group if they do not use it, and some believe it is annoying and inconvenient when one of their friends does not use IM. Most of the time, minors use these programs to communicate with real life friends; however some minors use public chat rooms and participate in one-on-one chats with complete strangers [2]. This gives predators the perfect environment to target underage victims and pursue them.

2.2 Similar Tools

There are commercial and network-level tools that can be used to protect children from Internet Predation. The most common alternative to SafeChat is a packet sniffer. Packet sniffers examine all the outgoing and ingoing traffic in a network and then apply a filter to only see the useful blocks of data. Many packet sniffers, and tools built upon their foundation, have easy to use UIs (User Interfaces). Parents, however, may have problems with these tools because:

1. They are too intrusive. Parents usually do not want to monitor all their children's chat data or be too intrusive into their children's social life.
2. They require parents to read through a lot of trivial chat data. If predation detection is not designed into a tool, the parent will need to read through the data by hand.
3. Tools that are currently available to detect predation are based on a simple keyword matching and not communication theory. This brings the accuracy of these tools into question [8].

SafeChat is designed to overcome the limitations of other tools and provides a better alternative than the tools that are currently available.

Chapter 3: SafeChat 1.0

The first version of SafeChat was stand-alone software. SafeChat 1.0 used the WinpCap library. WinpCap is a library that gives programmers high level control over the retrieval and transmission of packets in the Windows environment. This library is used by many widely used commercial products such as Wireshark [3].

SafeChat 1.0 was designed to work with AIM Instant Messaging because AIM has the largest market share among IM tools. AIM uses a protocol called Open System for Communication in Realtime (OSCAR). Despite the name, OSCAR is not an Open Source System. In 2008, documentation on the OSCAR protocol was released [4]. In 2010 the "Open AIM" initiative was limited to existing users and the documentation on OSCAR is no longer available. According to AOL, their core business strategy is to work with selected partners and "Open AIM" conflicted with that strategy [5].

Like AIM, many other chat clients did not have proper documentation. SafeChat, to be a successful, has to be compatible with many protocols. We therefore looked into other ways that our goals could be accomplished.

Chapter 4: Design of SafeChat 2.0

SafeChat 2.0 is the current version of SafeChat. It is a third party plugin for the Pidgin, an open source instant messaging system. It uses detection algorithms to classify chat participants as potential predators.

4.1 Pidgin

Pidgin is one of the most popular open source instant messaging systems. It works on any Windows or Unix-based environment and supports multiple protocols including AIM, MSN, ICQ, IRC, and Yahoo. Unsupported protocols, like Facebook Chat, can be used in Pidgin with the use of third party plugins [5].

There are multiple reasons for choosing the Pidgin platform. The primary reason is that we want SafeChat to be available to assist as many families as possible. Therefore, SafeChat needs to support as many IM protocols as possible. Second, SafeChat can take advantage of the development efforts of the Pidgin community. When new protocols are made or existing protocols are changed, the Pidgin community will update Pidgin. This allowed us to focus on the predation algorithms for SafeChat instead of on infrastructure issues.

4.2 Chat History System

For SafeChat to achieve its goal it needs to keep track of all of the user's interactions. This is achieved using a basic XML file system. The chat logs are stored in a file with the name of "<other chatter>.xml". The software keeps track of every chat post between the user and the other chatter. The post attributes maintained in the XML structure are the sender, the receiver, the message, and the classification of the post (this is used in the detection algorithm and is detailed in section 4.3).

4.3 A Rule-Based Approach to Detection

SafeChat uses a rule-based approach for detecting predators. This is done by classifying chat posts into categories. When adult chatters use certain categories they are identified as predators.

4.3.1 Categories

SafeChat uses four different categories when classifying posts. The posts are categorized when they match a pattern. The patterns are based on term types.

The first classification category is the exchange or discussion of ***personal information.*** It has been assigned a nominal value of 200. Personal Information includes, but is not limited to, location, name, phone number, and address. Likes and dislikes, when discussed in a non-sexual manner, are included also in this category. Talking about relationships, whether it is family or romantic, in a non-sexual way is

categorized as 200. Discussion of hobbies, movies, and sports is also categorized as personal information.

Predators use personal information exchange to build common ground with their victim and collect information about the victim's support system. This information is used to strengthen the predator's relationship with the victim and weaken other relationships that the victim has in place. Posts are categorized this way when:

- A post contains an approach noun (i.e car, hotel) and a relationship noun (boyfriend, date), but does not contain a personal information noun (i.e. age).
- A post contains either an action verb (i.e. think, do) or question word (when, who), and a personal information noun.
- A post contains either an approach verb (i.e. come, see), or an action verb and a relationship noun.
- A post contains an activities noun.

The second classification category is known as ***grooming*** and is given the nominal value of 600. Grooming is the use of sexual terminology in any context. The sexual terminology can be explicit like asking the victim about sexual experiences or implicit such as misspelling words to give them a sexual connotation (i.e. "cum" instead of "come"). Reframing, which is the redefinition of non-sexual content into sexual terms (i.e. "I can help you become a woman") would also belongs in this

category. The predator uses grooming terminology to make the victim comfortable with the use of sexual terms. Posts are classified as grooming when:

- A post contains a communicative desensitization word (i.e penis, sex).
- A post contains either an action verb or communicative desensitization verb (i.e. kiss) and communicative desensitization noun (i.e. orgasm).
- A post contains either a second person pronoun or question word and a communicative desensitization verb.
- A post contains a first person pronoun, second person pronoun, or action verb and a communicative desensitization adjective (i.e. naked).

The third classification category is ***approach*** and has the nominal value of 900. Approach is an attempt made by the predator to meet the victim outside of the electronic domain. This can be an attempt to speak with the victim on the phone, acquire a victim's phone number or address, or meet victim in person.

Approach includes instances of the predator trying to isolate the victim from their support network, including instances of the predator trying to get the location of the victim's friends in order to identify the physical location of the victim. Attempts by the predator to make the victim lie or conceal things are also classified as an approach. Posts are classified as approach when:

- A post contains a first person pronoun, a second person pronoun, or an approach noun, along with an approach verb (i.e. come, meet), and does not contain an information noun.

- A post contains an approach verb, an action verb, or an isolation adjective (i.e. lonely), along with a family noun (i.e. dad, divorce).
- A post contains an isolation adjective and a second person pronoun.

Posts that are not categorized are given the nominal value of 000 and usually appear innocent [7].

4.3.2 Accuracy

The accuracy of this rule-based system was tested to see how consistent it was with three people categorizing the same 10 chat logs. Accuracy is defined as:

$$m/N$$

Where m is the number of times the computer system categorizes the predator posts the same as the human and N is the total number of posts in a chat log. This system achieved an accuracy ranging from 51.95% to 83.90% and with an average of 68.11%.

This system performed best with the posts that were categorized as grooming. The computerized approach got an average accuracy of 42.22% when compared to three humans. This is because the language a predator uses to groom a child is far less varied than when the predator is attempting an approach or exchanging of information.

The system performs poorly when attempting to categorize a post as an exchange of personal information. The average accuracy was only 15.6%. The rule-based algorithm has trouble when specific topics are discussed, such as particular movies, because the system uses a limited keyword dictionary.

Approach is categorized with better accuracy than personal information. When the system misclassifies an approach, it is usually because terminology is vague. Approach terminology also varies widely amongst different predators. This category is unlike grooming, which the rule based system categorizes well, because grooming nearly always contains certain terms but the set of terms the predator can use in an approach is much larger. The average accuracy for approach is 33.66% [7].

Posts that have zero coded categories are categorized the best with an average accuracy of 84.28%. This shows that the rule-based system can weed out the posts that do not contain predatory behavior.

This rule-based approach provided strong results for detecting Internet Predation considering the complexity and variation of human communication.

4.3.3 A Predator's Threshold

Because it is preferred for SafeChat to wrongly flag something as predatory then to miss flagging a predator, it errs on the side of caution. If SafeChat detects that the other chat participant is an adult (outlined in section 4.4) and has a post flagged as

600 or 900 category, there is a good chance that he is a predator, and SafeChat should take necessary steps to protect the minor.

4.4 Age Function

The age of the chat participants is very important. If the chat participants do not include a minor and an adult, despite what the categorization rules say, the situation cannot be a case of sexual predation. To tell the age of a chat participant, we developed rules. The rules were discovered by closely analyzing chat data from Perverted Justice. Perverted Justice is an organization that catches Internet Predators by having volunteers pose as teens in chat rooms and respond to adults that approach them for a sexual relationship [6]. We examined how and when ages were exchange and derived the following rules:

- A two digit number, the age of the message sender, is in a post preceded by a post with the terms such as “asl”, “a/s/l”, “old”, or “age”.
- A two digit number, the age of the message sender, is in a post that also contains terms such as “age” or “old”.

4.4.1 Accuracy of the Age Function

Using a set of 420 chat logs from Perverted Justice, we tested the age function to see how accurately the function identified the predator’s actual age. When we ran

the age function on these logs, it didn't detect any age in 68 of the logs and misidentified the age in 169 of the chat logs. However, it correctly identified 183 of the ages. We examined where the age function was going astray.

When the age function misidentified the ages, it was mostly due to the fact that predators often lie about their ages. However, they often don't lie about the fact that they are an adult; they are more likely to round their age down a bit. Due to this, we were more concerned about the times no ages were identified. This can happen because sometimes ages are acquired via forum or Myspace profiles instead of being exchanged in the chat. We outline where the function could miss an age in Section 4.4.2.

Since predators are not deceitful that they are an adult, it is more important that the age function properly detects that the chat participant is 18 years or older. The age function properly classifies 326 out of 420 predators from our test data as adults.

4.4.2 Pitfalls of the Age Function

There are a few pitfalls that occur from the age function. The rules can pull an incorrect age when a chat progresses like this:

Predator: I have a 12 year old brother

This can prove catastrophic. If a predator is classified falsely as a minor, SafeChat will not take the necessary steps to neutralize the threat. The rules can also give a chat

participant an incorrect age. This is because of the limitations of the rule-based approach. You can miss an age with chat data that could consist of something like:

Victim: A/S/L?

Predator: 22 u?

Victim: 12

However, there are a few possible alternatives to increase the accuracy. This can be examining more than one post after a term such as “A/S/L”. However, you risk pulling non-relevant information when a chat looks like this:

Predator: A/S/L?

Victim: 15/F/US.

Predator: Have any siblings?

Victim: I have a brother, he is 12

Predators can also lie about their age, but, surprisingly, they rarely say that they are minors. Obviously, if this occurs there is no computerized method to detect deceit.

Chapter 5:

SafeChat Implementation

An adult who wishes to use SafeChat to protect a minor, would first follow the instructions ChatCoder.com (these are not yet posted but will be when SafeChat is released). First, they will be instructed to install Pidgin. They will then download the SafeChat dynamic-link library (DLL) and put it in their pidgin directory. An easy to use installer will eventually be made so the software can be used by people from all backgrounds. After the dll is installed, the plugin can be enabled in Pidgin and SafeChat will begin monitoring chat communication.

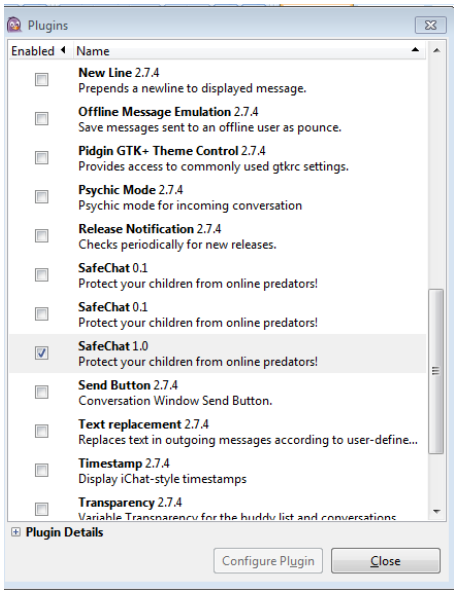

Figure 1: Enabling SafeChat

All chat activity will be logged into a directory in a file named as the other chat participant (shown in Figure 2).

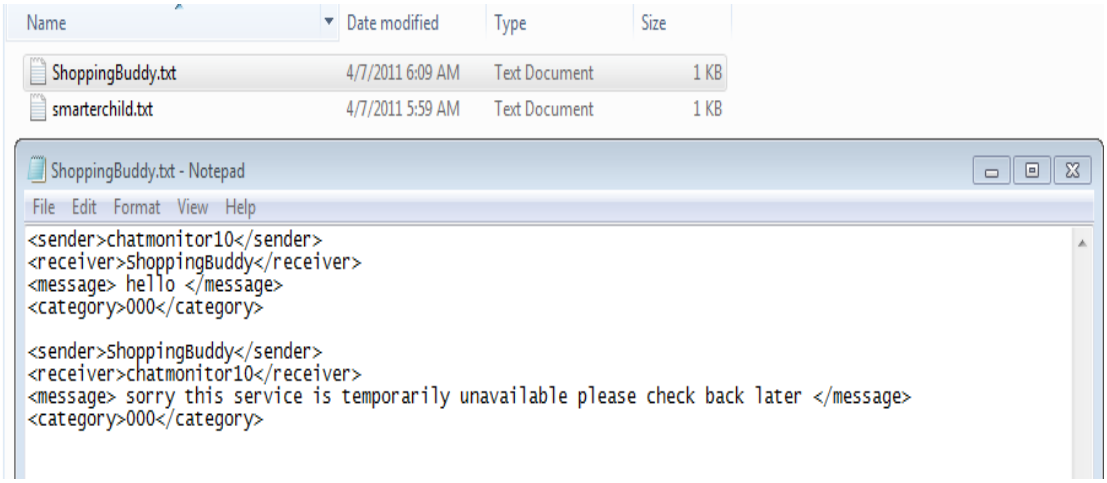

Figure 2: SafeChat's File System

Before the post is written to the file SafeChat runs the rule-based detection algorithm to categorize the post as a 000, 200 (Personal Information), 600 (Grooming), and 900 (Approach). To do this it must load the dictionary.

To increase the accuracy and make sure that no issues occur when the detection is run, some basic string manipulation is also performed.

1. SafeChat makes everything (dictionaries, post body) lowercase. This prevents having to search for “word”, “WORD”, “wOrd”. Usually the case of the letters adds nothing to the meaning of the post.

2. It converts netspeak to formal English. Instead of searching for “lol” and “laugh out loud”, the algorithm can just search for “laugh out loud”.
3. SafeChat adds spaces at the beginning and end of the dictionary terms and body. This prevents a search from returning a false positive. For example, searching for “word” against “sword” returns a positive despite a totally different meaning. But searching for “_word_” against “_sword_” returns a proper negative.
4. It strips the punctuation of the post’s body. This is to prevent terms from not being found correctly. For example, a search of “_word_” against “_word!_” will return a negative.

Once these manipulations are done, the post can be written to the proper file. Figure 3 shows an example of the post after the string manipulation.

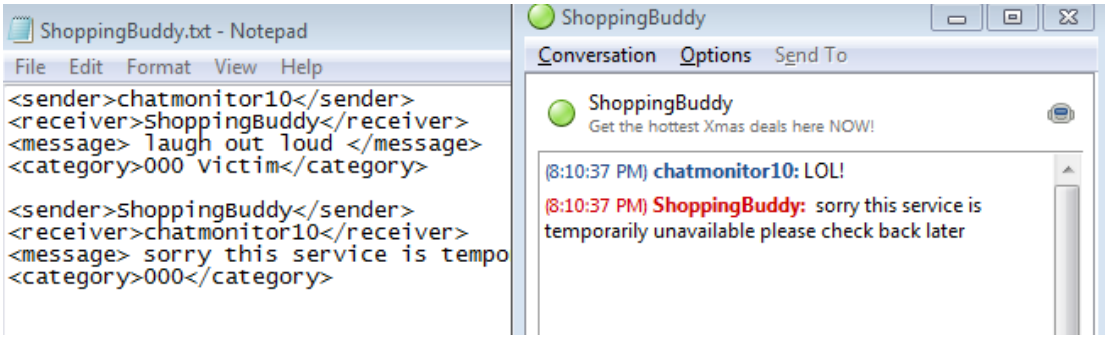

Figure 3: String Manipulations

Figure 4 shows a post being categorized as an approach.

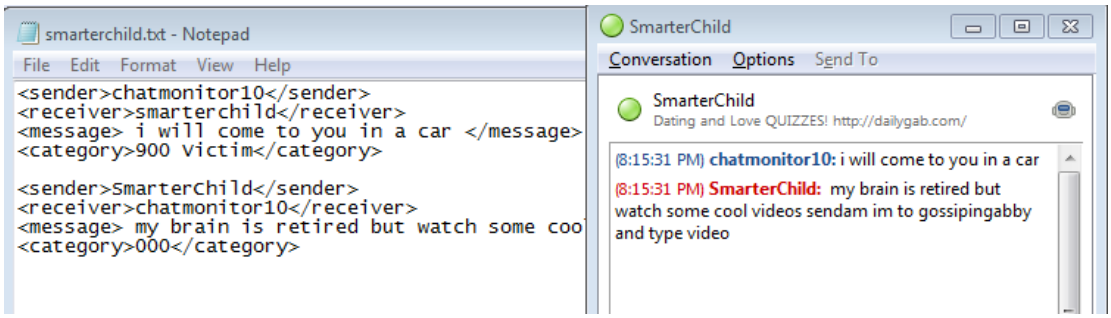

Figure 4: Categorized Post

SafeChat keeps track of all of the chat participants' ages in another xml file. This text file is checked periodically by the system. If an adult and a minor are chatting, there is a possibility that the adult is an Internet Predator.

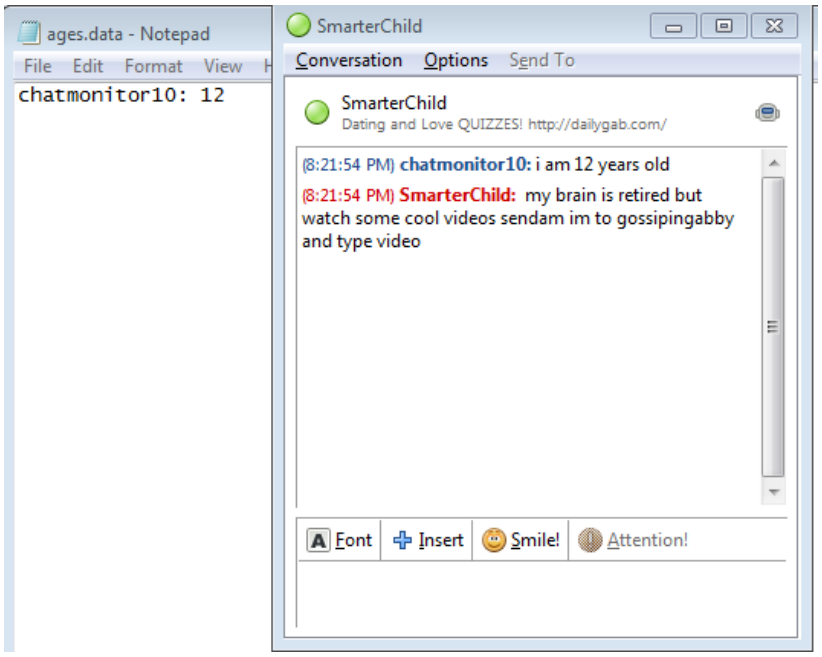

Figure 5: Age Detection

If SafeChat detects that a chat participant is an adult and has used a 600/900 post, it classifies the chatter as a predator. Before SafeChat is released, an email feature for parental notification will be added.

Chapter 6: Conclusion

In this thesis, we have detailed the design and implementation of SafeChat. SafeChat, a Pidgin plugin, is designed to detect cases of Internet Predation. It is superior to current chat-monitoring tools because it uses an age function and rule-based predation detection based on communication theory [7]. SafeChat allows parents to avoid reading through a lot of unimportant data, and allows them to still protect their children, while at the same time allowing their children to have privacy.

SafeChat detects an adult chatter 77% of the time and then runs a rule-based approach that categorizes the posts with 68% accuracy. SafeChat provides a strong foundation for future research on the detection of Internet Predators.

Several features need to be added before SafeChat can be released. Most importantly, it needs a mechanism for emailing parents or blocking predators. We also plan to add a feature that will collect chat data and send it to our project team for research purposes.

References

1. Wolak, Mitchell, and Finkelhor. Online Victimization of Youth: 5 Years Later. National Center for Missing and Exploited Children. 2006.
2. Grinter and Palen. Instant Messaging in Teen Life. Proceedings of the 2002 ACM conference on Computer supported cooperative work. 2002.http://www.winpcap.org
3. http://web.archive.org/web/20080308233204/http://dev.aol.com/aim/oscar/
4. http:/dev.aol.com/aim
5. http://pidgin.im
6. Perverted Justice. www.perverted-justice.com.
7. I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride and E. Jakubowski. Learning to Identify Internet Sexual Predation. In International Journal of Electronic Commerce. To appear.
8. Kontostathis, April, Lynne Edwards, and Amanda Leatherman. (2009). Text Mining and Cybercrime In Text Mining: Application and Theory. Michael W. Berry and Jacob Kogan, Eds., John Wiley & Sons, Ltd. 2009.