A method to calculate social networking hazard probability in definite time

Dimitrios Michalopoulos[1], Ioannis Mavridis[1]

[1] Department of Applied Informatics, University of Macedonia, 156 Egnatia Street 54006 Thessaloniki Greece
{dimich, mavridis}@uom.gr

Abstract

This work integrates with hazards for minor users while they are exposed to social networks. In particular, it contributes with the statistical relationship of these hazards with the exposure time as well as the amount of published personal information. Working on this direction, an experiment was conducted that has revealed a huge number of personal information exposed by users of social network applications. Moreover, a significant amount of suspicious activity against minors has been recorded. Experimental data led to the hypothesis that online hazards can be modeled with known statistical distributions. In order to examine this hypothesis, survival analysis techniques, which involve the estimation of certain functions that reflect the relation of a disastrous event with time, were applied. In particular, the distribution of the rate at which suspicious activities towards children occur in social networks, as they were recorded through the experiment, was derived. The results show that the incoming hazards for minor female profiles follow the Logistic distribution, while the corresponding hazards for minor male profiles follow the Normal distribution. Such knowledge is considered to be crucial for developing an effective system for automated grooming recognition in real time by optimizing the detection threshold as a function of time. Thus, the threshold sensitivity can be appropriately adjusted such that lower frequencies of occurrence lead to lower threshold sensitivities, and higher frequencies of occurrence lead to higher threshold sensitivities.

Keywords

Social networks, grooming, sexual exploitation, survival analysis, privacy leaks, distribution fitting, minors' hazards

1 Introduction

Modern communication media has brought many hazards for minor users. Along with the opportunity for intercultural communication, this evolution allows pedophiles and sexual Cyber-predators to attract their victims. Many children and teenagers have become victims of online sexual exploitation attempts (Armagh, *et al.* 2006). This phenomenon is generally known as grooming (O'Connell 2003). The consequences for grooming victims are catastrophic and many child victims are harmed for the rest of their lives (Berson 2003).

Child grooming occurs in every country, civilization, religion or ethnic group, and incidents are dramatically increasing. Cyber-predators are usually using social networks for communication, as well as searching and attracting new victims. According to experts, predators never before had the opportunity to communicate so directly with their victims as they do online (Olson *et al.* 2007). Specifically, using typical strategies attract their victims presenting themselves as experts to particular aspects of victim's interests, gain the necessary trust and then perform their attacks. The problem is getting worse as many of the victims are afraid to talk about these incidents with local authorities. Besides, social networks do not help authorities for limiting these phenomena. It is indicative that of 292 complaints for Facebook users submitted to Child Exploitation and Online Protection Centre (Ceop) in UK, none of them came directly from Facebook (Edwards 2010).

As the problem of online grooming is recent, there in not much published related research. In a similar work, Kontostathis *et al.* have analyzed the challenges of creating effective defenses against child sexual exploitation (Kontostathis *et al.* 2009). In addition Olson *et al.* has analyzed the strategies which sexual predators follow to achieve their goals (Olson *et al.* 2007). Similarly, predators' approaches are studied by O'Connell, revealing the nature of online grooming attacks (O'Connell 2003).

The research work presented in this paper was developed in the context of our general effort that is mainly focused on creating defenses against grooming attacks. In particular, this work researches on the hazards for minors as they are exposed on social Media and especially on Facebook. More specifically, we investigated the relation of the hazards in social networks with the time children are exposed to them (exposure time) as well as on the amount of personal information that is exchanged with strangers. To address this issue, we adopted techniques used in survival analysis. These techniques involve the estimation of certain functions which reflect the relation of a disastrous event with time. In particular, we initially extracted an experimental data set from Facebook by creating 10 profiles and collecting all data that indicate a potential risk (incoming friend requests, requests to date applications). We then noticed that hazards' occurrence rate varies with time and gender. Utilizing methods used in survival analysis, we made the hypothesis that incoming risks can be modeled for each gender by existing statistical distributions. Using proper tools, we calculated the parameters that optimize the distribution fitting, thereby testing the validity of our hypothesis. The verification of our hypothesis provides us with the ability to calculate the hazards' probability as a function of time and therefore create effective defenses.

This paper is structured as follows: Section 2 provides the framework in the context of which the work of this paper is developed. Section 3 describes the experiment scenario and corresponding results, whereas section 4 provides a brief description of survival analysis methods that are utilized. The exercise of various distribution fitting tests is presented in section 5. The obtained results and their exploitation are discussed in section 6 and the paper is concluded in section 7.

2 The GARS framework

Our current research work is focused on creating defenses against hazards for minor users. For this purpose, the Grooming Attack Recognition System (GARS) was introduced in (Michalopoulos *et al.* 2010), which has been designed to transparently monitor Internet communications with full respect to communication privacy. The main objective is to provide in real time warnings about a potential threat to the designated parent. GARS takes as an input captured text and user nickname from communications and responds with a risk value corresponding to the risk for sexual exploitation attack. Input text could be sent from clients installed on a desktop/laptop or a mobile device (Michalopoulos *et al.* 2012).

The main functionality of GARS consists of two inputs (captured text and user nickname), three processing units (document classification, personality recognition and history checking) and corresponding output controllers that calculate the particular weighted factors and the total computed risk. Document classification is performed by utilizing known attack patterns that were extracted from perverted-justice.com website (perverted-justice 2012). The selected classification algorithm is the Naïve-bayes as tests have shown that it is the most accurate as well as the fastest (Michalopoulos and Mavridis 2011). Similarly, the second process performs personality recognition also based on dialogs (also extracted from perverted-justice.com) by predators who have been accused by the justice. Besides, the third process performs analysis of historical records (viz. if the same user has been recorded previously as suspicious).

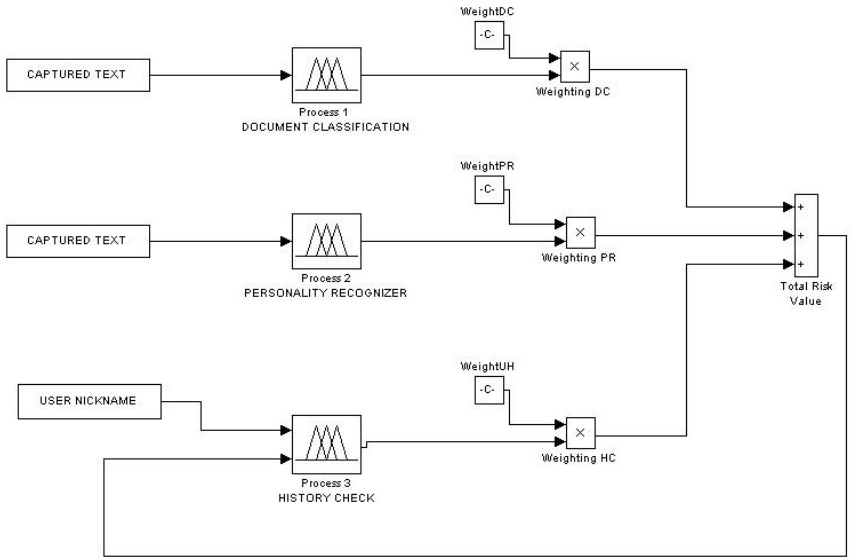

Figure 1. The GARS architecture

In every input, GARS responds with a weighted risk value that is recorded with the user's nickname to be used in the third process (viz. history check). Afterwards, according to decision thresholds, the decision making process results in sending a warning signal to the designated parent about the potential threat.

Definitely, the effectiveness of attack recognition depends on knowledge about hazards. Therefore, GARS demands for critical information about grooming hazards, such as severity, frequency and potential dependence with exposure time.

3 Experiment evaluation

The main purpose of the experiment was the identification of grooming hazards for minor users through social networks and the statistical relation with the exposure time. Besides, the experiment would provide more information about how teenagers are treated by other users and the amount of exchanged personal information. The experiment started with the creation of 10 Facebook profiles of minor users. Facebook was chosen as the most popular social network especially among teenagers. All profiles represented young teenagers aged 13 to 15 years old, half of them for male and the other half for female. Instead of clear face photographs, common images for youth were used. For example, boys used images for cars and famous sport teams, while girls used romantic images, actor and music group images. In all profiles, the default privacy settings (Facebook 2012), as well as the typical activities for a minor were entered. For example, joining groups of famous sport teams and music stars and registering in social and dating applications, like "Zoosk" and "Speed Dating". In addition, contact information (e-mail, IM) was adjusted to be visible to all other users.

Each profile was accessed daily. Besides, there were regular posts on each profile's wall, including common teenager thoughts, music videos or photographs of sports or music stars. All the above actions were necessary for the profiles to seem as ordinary teenager profiles.

The experiment lasted for 24 weeks. In the beginning of this period, all profiles were sending randomly to other profiles with common interests about 10 friend requests per day for the next 5 weeks (388 friend requests were totally sent). After that, all profiles kept going sending only 2 friend requests per week, according to Facebook's suggestions of mutual friends. Similarly, all incoming friend requests were accepted.

3.1 Data collection

All necessary test data were collected using the Facebook applications "Activity Statistics", "friendstats", "cha.fm", as well as the e-mail accounts connected with the profiles. Surprisingly, we discovered that the profiles had many friends and received friend requests and personal messages from many unknown so far profiles. Table 1 below presents statistics of the gathered results.

	Totalfriends	RequestsAccepted (of 388)	Percentage of Acceptance	RequestsAsked	Adults> 30	Adults 25-30	Adults 18-25	Minors<18
Female	1182	341	88 %	841	78 (7%)	147 (12%)	466 (39%)	491(42%)
Male	762	311	80 %	451	99 (13%)	149 (20%)	229 (30%)	285 (37%)

Table 1. Result statistics

As it is obvious from table 1, the summary of friends in the 5 female profiles is clearly higher than the corresponding summary of the 5 male profiles. Indeed, of the 388 friend requests sent from female profiles, 341 were accepted (viz 88%). Similarly, of the same (388) requests sent from male profiles 311 were accepted (viz 80%). Apparently, Facebook users are accustomed to accept friend requests from unknown profiles, while female profiles seem to be more acceptable than the male ones.

Similarly, the incoming friend requests where almost double in female profiles than in male profiles. Specifically, 841 profiles asked for friendship in female profiles during the 24 weeks of the experiment and 451 asked for friendship in male profiles. Therefore, female profiles seem to be more attractive to unknown users.

Although friendship activity is much higher in female profiles, the corresponding male profiles differ in the age of friend collection. In particular, table 1 denotes that 13% of friends in male profiles are aged over 30 years old and another 20% is aged 25-30 years old. In contrast, friends in female profiles are aged over 30 years old in percentage of 7% and respectively 12% are aged 25-30 years old. Therefore, female profiles attract more users from all ages, however male profiles attract more users in older (>25 years old) ages.

By the end of the experiment, the 10 Facebook profiles had gained access to huge amount of personal information through friendship integration with other users. Table 2 below presents the collected personal information.

	Albums	Photos	Videos	E-mails	Birthdays	Addresses
Female	2484	10185	1143	791	874	161
Male	1386	4712	781	497	681	76
Summary	3870	14897	1924	1288	1555	237

Table 2. Access to personal information

Results are indicating that Facebook users share a lot of their personal information with unknown users. However, such personal information can be easily exploited by third parties. For example, e-mail address can be used for massive spam or targeted advertizing considering information about favorite products, music or sports as it can

be extracted from relative Facebook groups. Besides, password guessing can be based on personal information like birthday or home address. Characteristically, most of the collected personal information was extracted from users under 25 years old and especially minors under 18 years old, although privacy settings in Facebook are stricter for underage users. Certainly, these results denote that users, and especially the younger ones, should use more wisely the privacy settings.

3.2 Suspicious activity

Having collected all test data after the experiment completion, we extracted all suspicious incoming activity for each one profile. As suspicious activity we consider all efforts or actions that can lead to child grooming. For example, a message with a link to inappropriate material, or a chat request with date intention. Specifically, as suspicious activity we reflect on personal messages – chat request, incoming friend requests, invitations in dating, posts in profile's Wall and any incoming activity from dating applications (like zoosk).

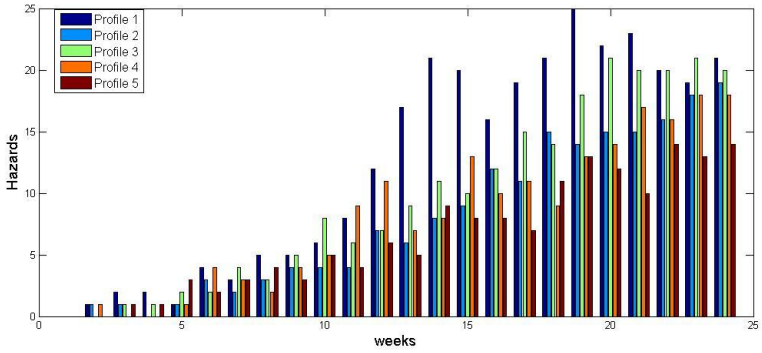

Figure 2. Suspicious activities per week for female profiles

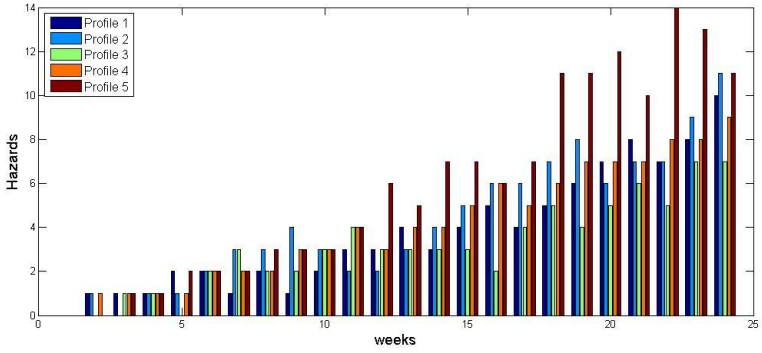

Figure 3. Suspicious activities per week for male profiles.

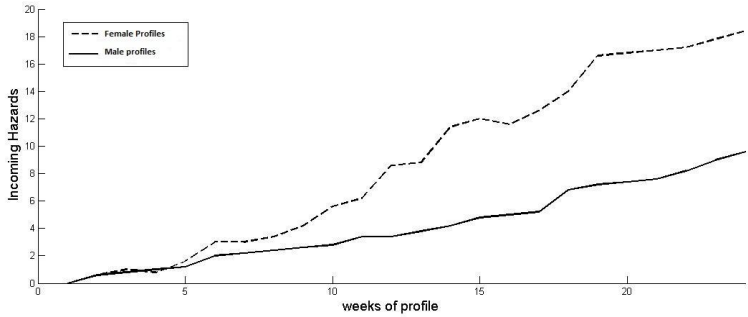

Figure 4. Plotting average of suspicious activities

Figures 2 and 3 present the collected data for suspicious activities, whereas Figure 4 depicts the evolution of the average of collected data for each gender as a function of time.

Indeed, the collection and categorization of suspicious activity was a challenging process. First of all, it was not automated by the usage of specific tools, but human recognition took place. For example, while profiles were receiving many personal messages, the majority of them were automated for commercial/advertizing purposes and few of them were considered as suspicious activity. Respectively, a large number of e-mail messages were delivered to designated e-mail addresses of profiles. The majority of these messages were generated by added Facebook applications; however there were some messages with surprisingly suspicious content. For instance, one of male profiles had been receiving e-mails from an unknown sender (without a clear user name in the email address) containing inappropriate content and defraud questions about sending more inappropriate material. The sender of these mails was probably someone who had access to the profile but would like to keep anonymity.

Summarizing, it is noticeable that female profiles received more suspicious messages however male profiles received less but more dangerous messages. The definition of dangerous messages reflects on messages that have clear malicious content from a person much older than the teenager. In contrast, female profiles also received malicious messages; however, the majority of them contained less malicious content. The next step was the identification of the statistical relation between these suspicious activities and the exposure time, as presented in the following section.

4 Survival Analysis

One of the major research tasks in health sciences is the identification of the risk factors for diseases, as for example the study on the connection between ionizing

radiation and leukemia (Le 1997). Such a connection can be verified by performing scientific investigation (Balakrishnan and Rao 2004). The usual steps for investigating the effects of an exposure to a risk factor are (Le 1997):

- Define the hypothesis proposal
- Investigate the hypothesis by testing or experiment
- Make a decision based on collected information, if the hypothesis is supported.

Survival analysis research includes studying groups of people with similar characteristics exposed to the same risk factor for a dedicated time period (David 2010). The basic aim of such a research is the identification of a potential statistical relation between the risk factor and the disease. Indeed, the important feature in such research is the time when the catastrophic event is going to happen. This time is commonly named as *survival time T.*

The distribution of the survival time *T* from the starting point until the catastrophic event is denoted by two functions: the *survival function S(t)* and the *hazard or risk function λ(t)*. The survival function *S(t)* is the probability that the patient survives longer than *t* time units (Le 1997). Therefore, if *T* is a continuous random variable and *F(t)* is the Cumulative Distribution Function (CDF) on $[0,+\infty)$, then it holds that (Papoulis and Pillai 2002):

$$S(t)=\Pr(T>t)=1-F(t) \tag{1}$$

The hazard function denotes the direct failure rate assuming the patient has survived to time *t* and is expressed as the probability of death in a time interval δ that tends to be zero (Le 1997):

$$\lambda(t)=\lim_{\delta\to 0}\frac{\Pr(t\le T\le t+\delta \mid t\le T)}{\delta} \tag{2}$$

For a small increment of δ, equation (2) yields (Klein and Moescberger 2003):

$$\lambda(t)=\frac{\left[-\frac{d}{dt}S(t)\right]}{S(t)} \tag{3}$$

Consequently, the formula (3) can be written as (Le 1997):

$$S(t)=e^{-\int_0^t \lambda(x)dx} \tag{4}$$

In health science (Miller *et al.* 1981), the estimation of survival and risk functions indicates the calculation of disease's spread in connection with time.. In addition, the estimation of the above functions is used for creating medicine treatments dedicated on the specific type of a disease and for comparing different treatments. For

example, when two different treatments are compared, researchers separate patients of the same disease into two groups, where the age composition of these groups is maintained as uniform as possible. The two treatments are implemented into the aforementioned groups such that, after a certain period of time, the survival and risk functions are calculated and compared. This comparison is used for identifying the most effective treatment (Le 1997).

Similarly to the survival analysis in health sciences, where the catastrophic event is death, we define the sexual exploitation of the minor user to be the catastrophic event in social network Internet communications. The risk factors where minor users are exposed are online hazards. Cyber-predators follow different strategies on approaching their victims and thereby performing their grooming attacks. Usually, they implement the so-called "hit and run" method, which refers to a vast attack against the minor user (O'Connell 2003). In other cases, they put into practice more sophisticated techniques by spending more time on knowing their victim and acquiring details of victim's personal life (O'Connell 2003). Therefore, the catastrophic event of child grooming may occur in an unexpectedly short period of child's exposure time.

This work aims at identifying the statistical relation between malicious approaches and the minor's exposure time (Papoulis and Pillai 2002). Similarly to the survival analysis in health sciences, where the estimation of the above functions can be used for improving the medication treatment, such calculation can be utilized for improving GARS effectiveness with variable detection thresholds.

5 Distribution fitting

Having collected the data set (suspicious activities) from the experimentation scenario, we made the hypothesis that hazards in social networks can be modeled with known statistical distributions. In case the hypothesis was true, we then could model incoming hazards in connection with time and therefore create effective defenses in the context of the GARS framework.

To verify the above hypothesis, we used the Matlab's distribution fitting tool (MathWorks Inc. 2005) and the Kolmogorov-Smirnov tests (Papoulis and Pillai 2002). The former identifies the known standard distributions which seem to be closer for fitting with the empirical distribution function of the sample data (captured through the experiment), whereas the latter compares the hypothesized (empirical distribution function of the sample data) with a reference probability distribution (Papoulis and Pillai 2002).

The known distributions used for fitting were the following: Normal (N), Generalized extreme value (GEV), Exponential (E), T location Scale (TLS), Logistic (L), Extreme value (EV), Generalized pareto (GP), Rayleigh (R), Gama (G) and Weilbull (W).

5.1 Average data fitting

Figures 5 and 6 represent the distribution fittings for average female data and average male data, respectively. The parameters for standard distributions GEV, E, N, TLS, L, EV ang GP were extracted from Matlab "dfiittool" (MathWorks Inc. 1999; MathWorks Inc. 2005). Similarly, using Matlab's standard functions ("raylfit", "gamfit" and "wblfit"), we calculated the parameters for distributions R, G and W (MathWorks Inc. 1999).

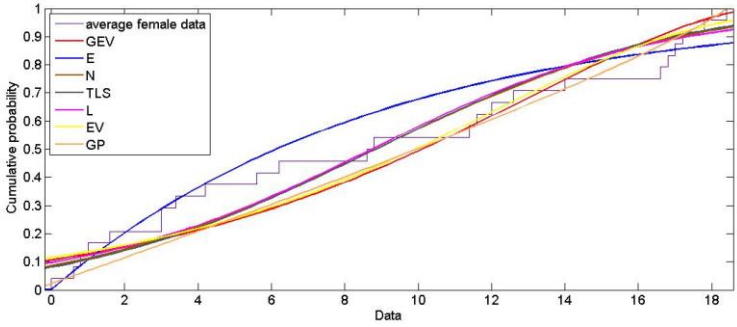

Figure 5. Distribution fitting for average female data

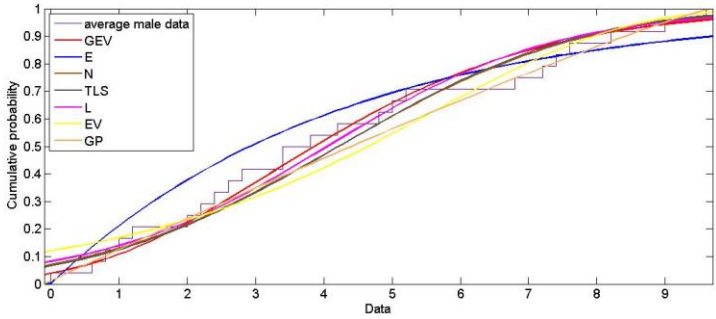

Figure 6. Distribution fitting for average male data

5.2 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test (KS-test) is used to assess the level of proximity between two data sets (Marsaglia 2003). In our case, we used the one-sample KS-test (MathWorks Inc. 2005), which calculates the distance *D* between the empirical distribution function of the sample data and a standard distribution function. Specifically, we used the one-sample KS-test as a "goodness of fit", viz. a statistical

model which describes how well a set of observations can fit a standard distribution (Massey 1951). The test calculates the distance *D* as:

$$D = \max(|F(x) - G(x)|) \tag{5}$$

Where $F(x)$ is the hypothesized (empirical) distribution function of the sample data function and $G(x)$ is a standard distribution.

We applied the KS-test on the average of both genders' data sets (female and male) and with various standard distributions. Parameters for standard distributions were extracted from "dfittool", "raylfit", "gamfit" and "wblfit" functions of Matlab. The alpha value used for all tests was *0.01*, instead of default *0.05*. Specifically, the alpha value represents the probability that the test fails if Matlab returns that hypothesis is true (Papoulis and Pillai 2002). Beyond the logical *h* (*h=1* if the hypothesis is rejected, while *h=0* if the hypothesis is true), we calculated the p-value *p* and the test statistic *k*. Assuming that the hypothesis is true, *p* is the probability of getting a test statistic at least as high as the one that was actually calculated (Stuart and Ord 1994; Bharath 2010). However, *p* is not the probability that the initial hypothesis is true (Marsaglia 2003). A statistic value on which the result is based on whether to accept or reject a hypothesis is the test statistic *k* (Bharath 2010). More specifically, *k* is the maximum difference between the curves, viz. the hypothesized and the standard one (Papoulis and Pillai 2002). Therefore, the criterion for a comparison between distributions is the lowest *k* value. Results for average female and male fitting tests are presented in tables 3 and 4.

Distr.	N	GEV	E	TLS	L	EV	GP	R	G	W
h	0	0	0	0	0	0	0	0	0	0
p	0.6918	0.5088	0.3546	0.6524	0.7186	0.5578	0.5426	0.1058	0.3384	0.3535
k	0.139	0.1613	0.1829	0.1437	0.1357	0.1552	0.1571	0.2401	0.1854	0.1831

Table 3. Average female fitting

Distr.	N	GEV	E	TLS	L	EV	GP	R	G	W
h	0	0	0	0	0	0	0	0	0	0
p	0.8877	0.7922	0.4479	0.8731	0.3314	0.6536	0.5205	0.5707	0.7691	0.4095
k	0.1127	0.1265	0.1694	0.1150	0.1866	0.1436	0.1598	0.1536	0.1295	0.1747

Table 4. Average male fitting

6 Discussion

Based on the fittings of tables 3 and 4, we conclude that the hypothesis of average experimental data fitting with standard CDFs is true for all distributions of both

female and male profiles' average of test data. Indeed, for concluding on which distribution can fit more accurately to the captured data set, we used as a criterion the lowest *k* value (Papoulis and Pillai 2002).

From table 3 it can be extracted that the average female data set (Table 3) best fits on the Logistic distribution (CDF) with μ = 8.73565 and σ = 3.92103. Therefore, formula 6 presents the distribution function for female hazards:

$$F_{female}(t) = \frac{1}{1+e^{-(t-8.73565)/3.92103}} \tag{6}$$

Similarly, the average male data set best fits (Table 4) best fits on the Normal distribution (CDF) with μ = 4.21667 and σ = 2.85287. Similarly, formula 7 presents the distribution function for male hazards:

$$F_{male}(t) = \frac{1}{2.85287\sqrt{2\pi}} \int_{-\infty}^{t} e^{-(x-4.21667)^2/16.277} dx = \frac{1}{2}[1+erf\left(\frac{t-4.21667}{2.85287\sqrt{2}}\right)] \tag{7}$$

Where *erf* is so-called *error function* (Spiegel *et al*. 2000).

The above results were extracted from the average data set for each gender. The verification about the satisfactorily fitting of each profile's data set with the above corresponding distributions was performed with 10 KS-tests (one for each data set). At this point the hypothesis was that the corresponding data set did not differ significantly from formula (6) for the female profiles and from formula (7) for male profiles. Table 5 below denotes that hypothesis is true for 9 out of 10 data sets. The hypothesis is rejected only for the first female profile.

Profile	Female 1	Female 2	Female 3	Female 4	Female 5
KS-Test	1	0	0	0	0
Profile	Male 1	Male 2	Male 3	Male 4	Male 5
KS-Test	0	0	0	0	0

Table 5. Applying KS tests to all profiles' data sets

In order to calculate the corresponding survival functions, formulas (6) and (7) yield from equation (1) as formulas (8) and (9).

$$S_{female}(t) = 1 - \frac{1}{1+e^{-(t-8.73565)/3.92103}} \tag{8}$$

$$S_{male}(t) = \frac{1}{2}[1 - erf\left(\frac{t - 4.21667}{2.85287\sqrt{2}}\right)] \tag{9}$$

Similarly, to calculate hazard functions, formulas (8) and (9) yield from equation (3) as formulas (10) and (11) (MathWorks Inc. 1999):

$$\lambda_{female}(t) = \frac{0.255035e^{0.255035t}}{9.32778 + e^{0.255035t}} \tag{10}$$

$$\lambda_{male}(t) = \frac{0.279678e^{-0.0614336(-4.21667+t)^2}}{1 - erf[-1.04514 + 0.247858t]} \tag{11}$$

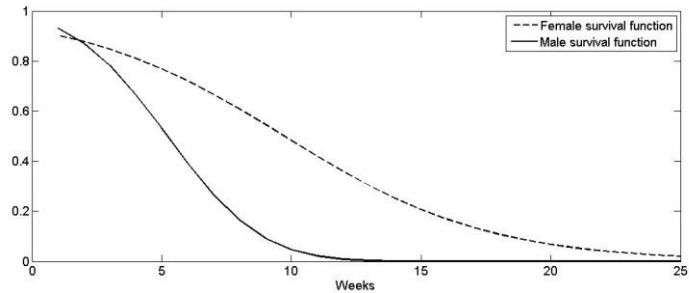

Figure 7. Survival Functions plot

Figures 7 and 8 depict the graphs of the calculated survival and hazard functions, respectively. It is indicative that even though incoming hazards for female profile are more in absolute numbers, the surge in the rate of occurrence in male hazards results in a sharper curve of male survival function. The sharpness in male hazards is more obvious in Figure 8. These results correspond to the observation that although male profiles have received less suspicious messages than females, those messages were more dangerous. Thus, grooming risk for males is estimated to be higher than the risk for females, as they are exposed online.

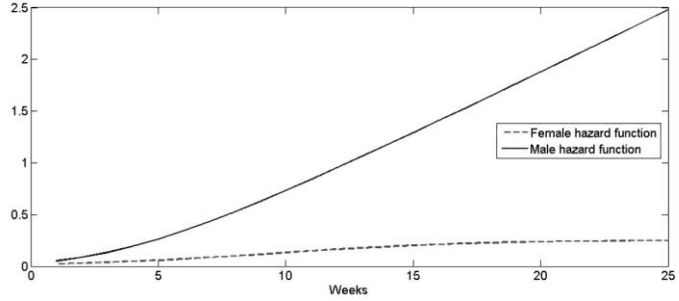

Figure 8. Hazard functions *λ(t)* plot

The above functions indicate the connection of children hazards with the exposure time. Specifically, based on the calculated CDF function (formulas 6 and 7) the probability of a hazard occurrence can be calculated accurately for a specific time period. As a result, it follows from the definition of the density function (CDF) that for any time instances t_a and t_b such that $t_b > t_a$, the hazard probability of drying time from t_a until t_b is obtained as (Papoulis and Pillai 2002):

$$p = F(t_b) - F(t_a) \tag{12}$$

Where *F(t)* denotes the CDF evaluated at time *t*. Let us now consider an infinitesimally small time interval, i.e., $t_b - t_a = \Delta t$. Then, (12) yields

$$\frac{F(t_b) - F(t_a)}{\Delta t} = \frac{p}{\Delta t} \tag{13}$$

When the value of Δt tends to 0, the left hand side of (13) equals the derivative of *F(t)*, i.e., the probability density function (PDF) *f(t)*. (Hijab 2007). Therefore, (13) yields:

$$p = f(t)\Delta t \tag{14}$$

In particular, PDFs can be calculated from corresponding CDF functions. Therefore, formulas (6) and (7) yield formulas (15) and (16):

$$f_{female}(t) = \frac{e^{-(t-8.73565)/3.92103}}{3.92103(1+e^{-(t-8.73565)/3.92103})^2} \tag{15}$$

$$f_{male}(t) = \frac{1}{2.85287\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{t-4.21667}{2.85287})^2} \quad (16)$$

Consequently, from formula 14 yields that for a sufficient small time difference Δt the hazard probability can be calculated from formulas (15) and (16) multiplied by Δt. Otherwise, the hazard probability is calculated from formula (12). Therefore, using the above formulas GARS can calculate hazard probability in definite time. Specifically, GARS can be adjusted to the specific needs of children profiles according to gender and exposure time. Current work concludes on formulas which indicate the variation of hazards as a function of time for both genders. Consequently, detection thresholds can be either more strict when the above formulas indicate higher grooming probability or less strict when lower grooming probability is calculated.

7 Conclusions

This work integrates with our overall research effort on creating defenses against grooming attacks, as impressed in the framework of GARS. Specifically, the experimentation which researches on hazards for minor users while they are online in social networks was presented. Experiment results have shown that users accept friend requests from unknown users and share with them personal information. For example, after the experiment duration time fake profiles had gained access to significant amount of personal information. This information can be manipulated for massive or spam message generation or profile's password guess using birthdays or home addresses. Moreover, female profiles are more attracted to younger ages (under 25 years old) where as male profiles are more attracted to older ages. Besides, although male profiles had less friendship interconnections, have received more dangerous requests in terms of sexual exploitation attacks.

In addition, the relation of minor users' hazards in social networks with the exposure time was researched. The obtained results demonstrated that online hazards for minor users follow specific statistical distributions for each gender. In particular, female profiles follow the *Logistic* distribution whereas male profiles follow the *Normal* distribution. Moreover, it is indicative that male profiles are exposed to greater grooming risk as they are exposed in social media. These conclusions are useful in predicting the incoming hazards for each new child registered profile. In particular, specific formulas have been conducted which calculate the hazard probability in designated time.

8 References

Armagh D. S. and Battaglia N. L. (2006). "*Use of computers in the sexual exploitation of children. Portable guides to investigating child abuse*."

Washington, DC, U.S. Dept. of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention.

Balakrishnan, N. and C. R. Rao (2004). "*Advances in survival analysis*". Amsterdam ; Boston, Elsevier.

Berson, I. (2003). "*Grooming Cybervictims: The Psychosocial Effects of Online Exploitation for Youth*." Journal of School Violence **2**(1).

Bharath, R. (2010). "*Nonparametric statistics for non-statisticians: a step-by-step approach.*" Choice: Current Reviews for Academic Libraries **47**(7): 1324-1324.

David, A. P. (2010). *Survival analysis using SAS a practical guide*, second edition. Cary, N.C., SAS Pub.

Edwards, R. (2010). "*Complaints about grooming and bullying on Facebook quadruple*" available at: http://www.telegraph.co.uk/news/uknews/crime/7567922/Complaints-about-grooming-and-bullying-on-Facebook-quadruple.html (accessed 10 July, 2012).

Facebook. (2012). "*Data use policy*" available at: http://www.facebook.com/about/privacy/ (accessed 23 December, 2011)

Hijab, O. (2007). *Introduction to calculus and classical analysis.* second ed. New York, Springer.

Perverted Justice *"Perverted-Justice.com Perverted Justice."* available at: www.perverted-justice.com (accessed 27 September 2010)

Kontostathis, A, Lynne E. and Leatherman A. (2009). "*Text Mining and Cybercrime*" In Text Mining: Application and Theory. Michael W. Berry and Jacob Kogan, Eds., John Wiley & Sons, Ltd. 2009.

Le, C. T. (1997). *Applied survival analysis*. New York, Wiley.

Marsaglia, G., W. Tsang, and J. Wang (2003). *"Evaluating Kolmogorov's Distribution.*" Journal of Statistical Software **8**(18).

Massey, F. J. (1951). "*The Kolmogorov-Smirnov Test for Goodness of Fit.*" Journal of the American Statistical Association **46**(253): 68-78.

MathWorks Inc. (1999). *MATLAB : the language of technical computing*. Natick, MA, MathWorks.

MathWorks Inc. (2005). *Simulink® : simulation and model-based design : using Simulink*. Natick, MA, MathWorks.

Michalopoulos D., Papadopoulos E. and Mavridis I. (2012). "*Artemis: Protection from Sexual Exploitation Attacks via SMS*". To be appeard in PCI 2012, Athens Greece.

Michalopoulos D. and Mavridis I. (2011). "*Utilizing Document Classification for Grooming Attack Recognition"*. The sixteenth IEEE symposium on Computers and Communications (ISCC'11) Kerkyra Greece

Michalopoulos D., Mavridis I. and Vitsas V. (2010). "*Towards a Risk Management Based Approach for Protecting Internet Conversations*". 9th European Conference on Information Warfare and Security, ECIW 2010**:** 201-208 Thessaloniki Greece.

Miller R. and Gong G. (1981). *Survival analysis*. New York, Wiley.

O'Connell, R. (2003) "*A typology of child cybersexploitation and online grooming practices* " Cyberspace Research Unit, University of Central Lancashire.

Olson, L. N., Daggs, J. L., Ellevold, B. L. and Rogers, T. K. K. (2007). "*Entrapping the Innocent: Toward a Theory of Child Sexual Predators Luring Communication*." Communication Theory **17**: 231-251.

Papoulis, A. and S. U. Pillai (2002). *Probability, random variables, and stochastic processes*. Boston, McGraw-Hill.

Spiegel, M. Schiller J. and Srinivasan R. (2000). *Schaum's outline of theory and problems of probability and statistics*. Schaum's outline series. New York, McGraw-Hill.

Stuart, A., Ord K. (1994). *Kendall's Advanced theory of statistics*. London, Edward Arnold.