Utilizing Survival Analysis for Modeling Child Hazards of Social Networking

D. Michalopoulos[1] and I. Mavridis[1]

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

Abstract

Social networks induce several hazards to children, which are correlated with the amount of time that children are exposed to those networks. To this end, this work investigates the relation of the aforementioned hazards with the exposure time. To address this issue, we adopt 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 derive the distribution of the rate at which suspicious activities towards children occur in social networks. This is accomplished through experiments on data sets extracted from Facebook. 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. This knowledge is then utilized for developing an effective system for automated grooming recognition, 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

1 Introduction

The evolution of communication media leads to new forms of experience in our days. Social networks are becoming very popular among teenagers, mainly through Facebook, which is reaching almost 800 million active registered users (Facebook 2012). However, these new forms of communication have raised significant hazards for minor users. Many children and teenagers have become victims of online sexual exploitation attempts (Armagh, Battaglia *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 2007).

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. For this purpose, the Grooming Attack Recognition System (GARS) has been introduced in (Michalopoulos *et al.* 2010), which is designed to transparently monitor internet communications with full respect in communication privacy. Moreover, we have published research work on privacy and security leaks of social networks (Michalopoulos and Mavridis 2010). This work has revealed that Facebook users tend to make public their personal data and exchange personal information with strangers. 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 2009). In addition Olson et al. has analyzed the strategies which sexual predators follow to achieve their goals (Olson 2007). Similarly, predators' approaches are studied by O'Connell, revealing the nature of online grooming attacks (O'Connell 2003).

More specifically, we investigated in this work the relation of the hazards in social networks with the time children are exposed to them (exposure time). 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. More specifically, we initially extracted an experimental data set from Facebook by creating 10 fake 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' rate of occurrence as a function of time, for each gender. Subsequently, this knowledge can be used for optimizing the detection threshold (viz. in GARS) as a function of time. For example, the threshold sensitivity can be adjusted such that lower frequencies of occurrence lead to lower threshold sensitivities, and higher frequencies to higher threshold sensitivities.

This paper is structured as follows: Section 2 provides a brief description of survival analysis methods that are utilized in our experimentation scenario described in Section 3. The exercise of various distribution fitting tests is presented in Section 4, and the obtained results are discussed in Section 5. Our conclusions and future work are included in Section 6.

2 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 studding of groups of people with similar characteristics exposed to the same risk factor for a dedicated time period (Allison and Books24x7 Inc. 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 happened. 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 is expressed as 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 it is the probability of death in a very small time interval δ (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, Gong *et al.* 1981), the estimation of the survival function and the risk function indicates the calculation of disease's spread as a function of time. The above functions are used for creating medicine treatments dedicated on the specific type of a disease. In addition, the estimation of the above functions of survival analysis can contribute to the comparison of 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 Internet user to be the catastrophic event in our research work. 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 learning details for victim's personal life (O'Connell 2003). Therefore, sometimes the catastrophic event of child grooming occurs in a short period of child's exposure time, while other times this event occurs after a longer period of time.

The main purpose of this work is to identify 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 used in our research for improving GARS effectiveness with variable detection thresholds.

3 Experimentation scenario

In order to identify the relation between grooming hazards and the exposure time, we made the hypothesis that incoming hazards in social networks follow statistical distributions. In order to collect a data set of hazards related to the time unit, we created 10 Facebook profiles of minor users. As time unit we assumed that of one week and the whole experiment lasted for 24 weeks.

All profiles represent 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 we used the default privacy settings (Facebook 2012), as well as the typical activities for a minor. For example, joining groups of famous sport teams and music stars and registering in social and dating applications, like "Zoosk" and "Speed Dating".

The whole experiment lasted for 24 weeks. In the beginning of this period, our profiles were sending randomly to other profiles with common interests about 10 friend requests per day for the next 5 weeks. After that, our profiles kept going sending only 2 friend requests per week, according to Facebook's suggestions of mutual friends. By the end of experimental time, we collected for further analysis all necessary test data using the Facebook applications "Activity Statistics", "friendstats", "cha.fm" as well as the e-mail accounts connected with our profiles. Surprisingly, we discovered that our profiles had many friends and received friend

requests and personal messages from many unknown so far profiles. Summarizing, we collected all suspicious incoming activity for each one profile. As suspicious activity we consider each activity 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).

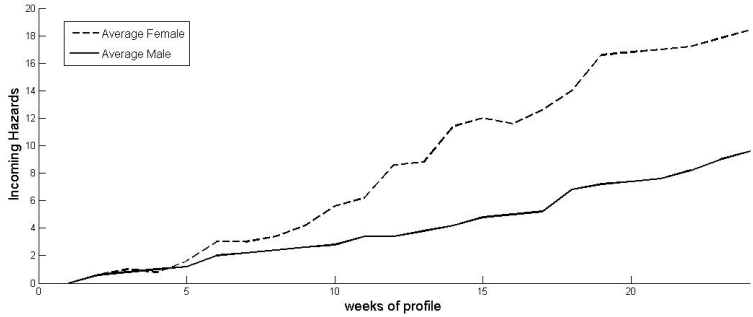

Figure 1: Plotting average incoming hazards

Figure 1 depicts the evolution of the average collected data for each gender as a function of time.

4 Distribution fitting

Having collected the data set from our experimentation scenario, we made the hypothesis that hazards in social networks can be modeled with known statistical distributions. If this hypothesis is true, we can model incoming hazards as function of time and therefore create effective defenses. To verify this 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 distributions which are closer for fitting with the hypothesized which is created by the captured data set, whereas the latter compares the hypothesized with the existing distribution and concludes about their fitting.

The distributions we used for fitting were: 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).

4.1 Average data fitting

Figures 2 and 3 represent the female and the male average distribution fittings, respectively. The parameters for standard distributions were extracted from Matlab "dfiittool" for the first seven distributions (MathWorks Inc. 1999; MathWorks Inc. 2005). Similarly, using Matlab's standard functions ("raylfit", "gamfit" and "wblfit"), we calculated the parameters for distributions Rayleigh, Gama and

Weilbull (MathWorks Inc. 1999). These last three functions were not calculated with "dfittool" but with the corresponding Matlab's standard functions.

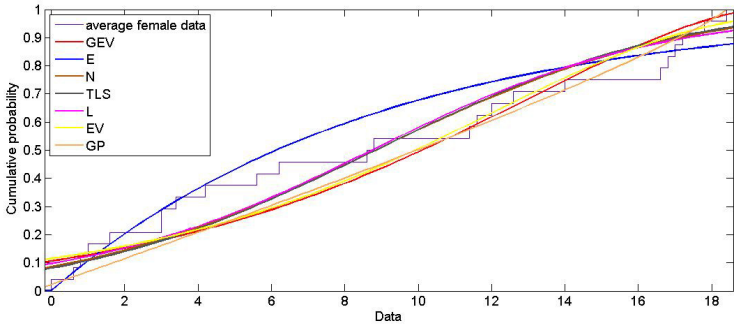

Figure 2: Distribution fitting for average female data with 6 distributions

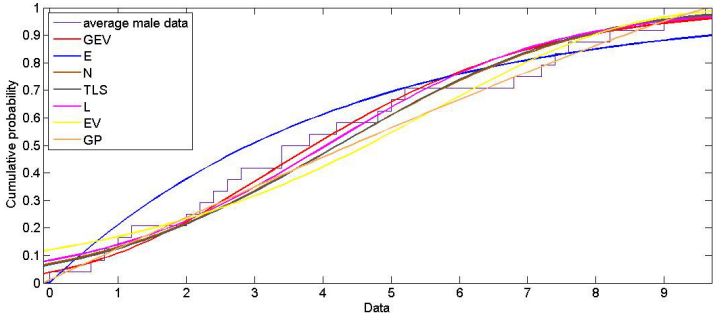

Figure 3: Distribution fitting for average male data with 6 distributions

4.2 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test (KS-test) is used to assess the level of proximity between two datasets (Marsaglia 2003). In our case, we used the one-sample KS-test (MathWorks Inc. 2005), which calculates a distance D between the empirical distribution function of the sample data and the standard distribution function. Moreover, 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 function and $G(x)$ is the standard distribution.

We applied the KS-test into the average of the test data for both genders' data sets (female and male) and with standard distributions. Parameters for standard

distributions were extracted from "dfittool", "raylfit", "gamfit" and "wblfit" Matlab's functions. The alpha value that used for all tests was *0.01* instead of default *0.05*. This alpha value represents the probability that the test fails if Matlab returns that hypothesis is true (Papoulis and Pillai 2002). The other values that we have calculated beyond the logical *h* (*h=1* the hypothesis is rejected, *h=0* the hypothesis is true) is 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, Ord *et al.* 1994; Bharath 2010). However, *p* is not the probability that the null hypothesis is true (Marsaglia 2003). A statistic value on which the result is based about 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 (Papoulis and Pillai 2002). Therefore, the criterion for comparison between distributions is the lowest *k* value. Results for female and male fitting tests are presented in tables 1 and 2.

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 1: 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 2: Average Male fitting

5 Discussion on results

Based on the fitting that are presented in tables 1 and 2, we conclude that the hypothesis of average experimental data CDF 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 1 it can be extracted that female data set best fits on the Logistic distribution (CDF) with $\mu = 8.73565$ and $\sigma = 3.92103$.

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

Similarly, male data set best fits (Table 2) on the Normal distribution (CDF) with $\mu = 4.21667$ and $\sigma = 2.85287$.

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

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

The above results were extracted from the average data set. In order to verify that the data set of each profile satisfactorily fits the above corresponding distributions, we used again the KS-test. At this point the hypothesis was that the corresponding data set did not differ significantly with (6) for the female profiles and with (7) for male profiles. Table 3 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 3: Implementing KS tests in all profiles' data

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):

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

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

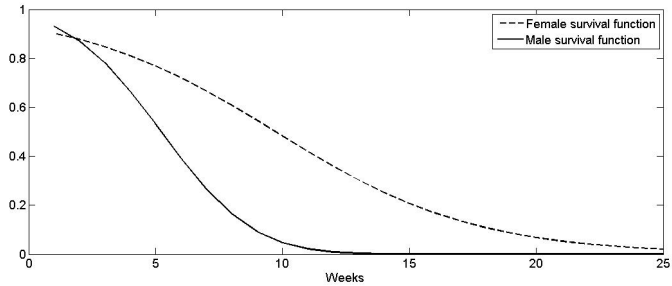

Figure 4: Survival Functions plot

Figures 4 and 5 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 sharper curve in male survival function. This is more obvious in Figure 5.

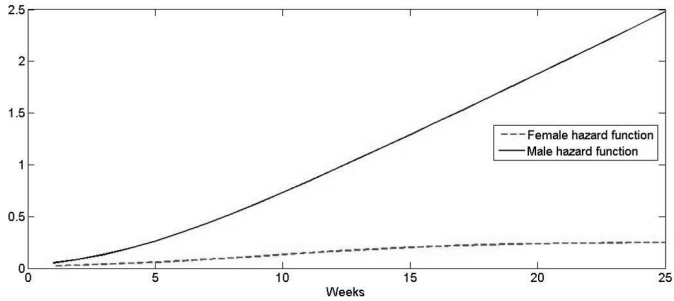

Figure 5: Hazard functions *λ(t)* plot

6 Conclusions

In this work we studied the relation of minor users' hazards in social networks with the exposure time. Our results demonstrated that online hazards for minor users follow specific statistical distributions for each gender. In particular, results shown that female profiles follow the *Logistic* distribution whereas male profiles follow the *Normal* distribution. These results provide us the opportunity to predict the incoming hazards for each new child registered profile. Such a statistical prediction can be very useful on creating defenses against online hazards. Moreover, the presented research work reveals that incoming hazards for children are increasing as a function of the exposed time. Therefore, when developing detection mechanisms the sensitivity of the detection threshold should be adjusted on the curves depicted above for optimum performance purposes.

7 References

"Child Sexual Abuse." Retrieved 9 November 2011, from http://www.nlm.nih.gov/medlineplus/childsexualabuse.html.

Allison, P. D. and Books24x7 Inc. (2010). *Survival analysis using SAS a practical guide*, second edition. Cary, N.C., SAS Pub.

Armagh, D. S., N. L. Battaglia, et al. (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.

Facebook. (2012). "Data use policy." Retrieved 23 December, 2011, from http://www.facebook.com/about/privacy/.

Facebook. (2012). "Statistics." Retrieved 6 January, 2012, from http://www.facebook.com/press/info.php?statistics.

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

Le, C. T. (1997). *Applied survival analysis*. New York, Wiley. **ISBN 0-471-17085-2**

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., 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.

Michalopoulos D. and Mavridis I. (2010). Surveying privacy leaks through online social networks *14th Panhellenic Conference on Informatics, PCI 2010*. **art. no. 5600443:** 184-187.

Miller, R. G., G. Gong, et al. (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. R., R. A. Srinivasan, et al. (2000). Schaum's outline of theory and problems of probability and statistics. *Schaum's outline series*. New York, McGraw-Hill.

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