A Probabilistic Model for Cyber Attacks and Protection

by

S.M. Asmoredjo

June 2005

Table of Content

Table of Figures

Table of Tables

Abstract

In this thesis an effort was made to model problems from the Information Security realm. In this age in which the transformation to the information age is well on its way the problems that arise due to the increased interconnectivity of the world seem to grow correspondingly. Actually, Information Security is not a new problem per se. But it is only that with the present fast developments that the issues have become increasingly more complex, and so, more sophisticated analysis have to be made. Already in ancient times it was found that battles could be won on tactical or strategic decisions. The basis for these decisions is information about for example enemy positions, supply trains, and hierarchy that can be exploited. Nowadays, armies still heavily depend on reconnaissance, but due to the information technology revolution the time span has decreased considerably. More examples of these dependency on information technology are public service area's including banking, transportation, and communication.
As a result, the problem of the dependence on technology becomes striking. It can safely be said that the absence of information technology will impact every day live dramatically. And as a consequence, excellent information security becomes more and more essential.

Due to the increased interconnectivity, the protection of information systems, particularly personal computers, has become a major topic. Unfortunately most software is not designed with the security of the computers in mind, and, as a result, frequently remains exploitable by malicious code. Furthermore, the only related security measures that are available at the moment use signature based virus detection systems and are unable to transcend the specifics to more general security detection systems. Detectors with low specificity are being developed but until now they have proven to have poor discrimination between valid code and malicious code. As a result of the lack of generality security programs need to be updated regularly. In the long run, however, this does not seem to be a proper solution as infection rates are increasing rapidly.
To improve awareness on the contributing factors to this problem an operational research approach was applied. The problem was addressed as a reliability problem and a Poisson process was used to model attack patterns. This model was chosen because of the strong ability to cope with changes such as the addition of parameters, and because of the apparent appropriate fit. This lead to the conclusive result that patch-times have a major impact on the capability of a virus detection system to assure security.

Another complex problem in information security is the question on the necessity to implement countermeasures. In current businesses decision making project return on investment calculations are the norm. Since risk mitigation is such an important component of security initiatives, these have to be framed in the return on investment of information security projects. And as a result the value of change of risk has to be incorporated in the analysis. Estimation of the value of change of risk however remains complicated and is not straigforward.
On the other hand, increasingly more enterprise are becoming aware of the impact of information security on business values. Often mission values include integrity and confidentiality. In some countries there is even the tendency to change legislation to force companies to show due diligence on information security related issues. As a consequence information security is getting high priority. In such cases, not only return on investment considerations are important but other factors such as legal need as well.

Often the choices of countermeasures and policies are extensive. A hazard minimizing approach is suggested as a counterpart for the return on investment strategy. It offers a different point of view on information security decision making. The model was able to show that factors such as the implementation length of projects as well as relative hazards could be of significance influence on information security related subjects.

Covariates play an important role in several quantitative models. One of the best known examples is probably the Cox proportional hazard model. This model is extensively discussed in literature, and also has captured my attention. For negatively correlated covariates it is shown that omission of one of them, results in a bias of the estimate for the other covariate. This bias is proven to be towards the null and depends on the correlation coefficient.

All in all, I enjoyed applying mathematics in the Information Security field where quantitative results are scarce. The advantage of the lack of quantitative results being that one is free to start anywhere, and the disadvantage that one does not know where to start. So, as a result the models are virtually standalone.
It was great to get the opportunity to work in a field in which so many contemporary, compelling problems are found. And I found it amazing how a simple, general model could still give detailed results which coincided with reality.

Acknowledgements

First of all I would like to thank the committee that has made it possible for me to write my thesis. The project supervisor T.A. Mazzuchi, who is a visiting professor at the TU Delft, professor at the George Washington University (GWU) and chairs the Engineering Management and System Engineering department at the GWU. Who has helped me greatly with the practical modeling.
Professor R.M. Cooke, who is the supervisor of the Risk and Environmental Modeling program in Delft and chairs the Applied Decision Theory department of Applied Mathematics at TU Delft. He has been a great inspiration and is a great mathematician and was a wonderful sounding board, especially in the early development of my thesis.
Professor J.J.C.H. Ryan, who is an assistant professor at the EMSE department of GWU and is the academic lead for the Information Security Management program at EMSE. Who has been a great aid in helping me to understand the vast area of information security and helped point out so many crucial details that became a significant part of my thesis.
And professor D. Kurowicka, who is an associate professor at the Probability, Risk and Statistics department at TU Delft. Who has been of great value in helping me out with the more practical aspects of graduating.

Also I would like to thank Professor D.J. Ryan, who is professor of Systems Management at the National Defense University for the remarks regarding the Information Security aspects that helped me improve this thesis.
As well as the help from fellow student C.J.L. Rothkrantz at TU Delft, who read the early draft versions of this thesis and helped me improve it a great deal.

Last but certainly not least, all people that support and have supported me during my studies. Thanks, for all that!

April, 2005

Martijn Asmoredjo

Outline

The thesis is subdivided into five chapters. The first chapter is designed to introduce the information security topic. Since every field has its own terms and terminology it was chosen to start of with the information security introduction such that in later chapters the terms and terminology could be used freely.
In the second chapter Poisson processes and their properties are introduced as they form the foundation on which the model is based. The Poisson process was chosen because its ability to model both the details and generalities well without forcing the structure of the model to change dramatically.
The third chapter discusses two examples in which the Poisson process is applied. The first example considers a virus as a particular cyber attack, this to illustrate and to emphasize the effect of patch-times. The second example considers the confidentiality aspect of information security and a hazard minimization approach for the implementation order of different safeguards is applied.
Chapter 4 addresses typical difficulties that arise in reliability data such as the presence of censored data. In practice, the Cox Proportional Hazard Model is widely used to handle censored data and is as such tested on characteristic data.
Finally, in the fifth chapter theoretical problems of the use and misuse of the Cox Proportional Hazard model are investigated. To this extend models in which the number of pertinent covariates are different are compared and lead to conclude that the choice of the number of covariates might bias the estimations of those coefficients, which was already known for independent covariates. It was found that for negatively correlated covariates, the dropping of one of the covariates leads to bias towards the null.

1 Introduction to Information Security

This chapter is designed to introduce the novice reader into the topic of information security. In no way it has been intended to capture all the various corners of the matter, this due to the sheer magnitude of the area. It is intended to make one obtain insight into the topic and for this the methodology of divide and conquer has been used. The information security topics that are elaborated upon will be used in the chapters where models are developed and tested. The goal of this chapter was to give a general overview of the material and to refrain from using a too technical language. Accordingly, do not expect to find configurations for firewalls or virus scans for as time spent would be in waste. But most of all, this chapter tells the tale of my pursuit to understand information security as for the start of my thesis, I too was a novice on this terrain.

Section 1.1 explaines what the information security concept is, and then in section 1.2 the information characteristics are treated. In section 1.3 examples out of everyday life are given to indicate how information security affects a wide range of individuals. In section 1.4 the information states are treated and in section 1.5 different threats are pinpointed. In paragraph 1.6 controls and countermeasures are defined and in section 1.7 the topic of phasing security is addressed.
All in all it is hoped that this chapter may be able to sketch the field of information security.

1.1 What is Information Security?

At the start of my thesis I had no idea what information security was, and, well frankly it took me some time getting accustomed to the notion since it is very global. The first idea was to consult a dictionary to look the two words up separately and than to combine the results. And what is a better place to look for creating a probabilistic model for cyber attacks than the World Wide Web.
So, 'googling' once on 'dictionary' resulted in the hit for www.dictionary.com, and yes I know that to Google still is no verb according to the linguists from the site, but to my opinion it should be in there. Is it not one of the most heard answers to a question the question asked in return, have you tried using Google?
Anyways, there were seven entries found on www.dictionary.com for information:

1. Knowledge derived from study, experience, or instruction.
2. Knowledge of specific events or situations that has been gathered or received by communication; intelligence or news. See Synonyms at knowledge.
3. A collection of facts or data: *statistical information.*
4. The act of informing or the condition of being informed; communication of knowledge: *Safety instructions are provided for the information of our passengers.*
5. *Computer science.* Processed, stored, or transmitted data.
6. A numerical measure of the uncertainty of an experimental outcome.
7. *Law.* A formal accusation of a crime made by a public officer rather than by grand jury indictment.

This explains some of the confusion since the word information alone already can be interpreted in seven different ways.
The fifth entry seems promising since it states that this interpretation is used in computer science but unfortunately I think it is rather to be thought of as possible states of information and not information on its own.
Furthermore, an enquiry for security resulted in another six possible interpretations:

1. Freedom from risk or danger; safety.
2. Freedom from doubt, anxiety, or fear; confidence.
3. Something that gives or assures safety, as:
    a. A group or department of private guards: *Call building security if a visitor acts suspicious.*
    b. Measures adopted by a government to prevent espionage, sabotage, or attack.
    c. Measures adopted, as by a business or homeowner, to prevent a crime such as burglary or assault: *Security was lax at the firm's smaller plant.*
    d. Measures adopted to prevent escape: *Security in the prison is very tight.*
4. Something deposited or given as assurance of the fulfillment of an obligation; a pledge.
5. One who undertakes to fulfill the obligation of another; a surety.
6. A document indicating ownership or creditorship; a stock certificate or bond.

Now the confusion is complete. Were there already seven interpretations for information the dictionary now came up with an additional six definitions for security. And as basic probability theory tells that the number of combinations in which the words could be used is the product of the separate interpretations, which results in 42 possible single interpretations!
But how too choose the right interpretation from all this possibilities?
I decided to read published works on the matter because it was said that there are a lot of papers and books on this topic, and surely they must contain a definition for the topic. As it turned out, there is indeed a vast amount of literature, but almost every one of them uses slightly different definitions.
After reading a lot of them I established that the quintessence from these definitions might well be that *information security* is about the protection of information and the systems that store, process and transmit the information.
Going back to the particular definitions of the words, with the aid of hindsight one can say that the third definition comes close but still needs a slight addition. So that information in this context is *a collection of facts or data put into context*.
I added the adjustment because to my opinion encrypted data also bears information but that without the proper decryption key it is nothing but gibberish.
And going back to the definitions from the dictionary, for security in the information security context I would be drawn towards the third entry *something that gives or assures safety* that seems reasonably vague enough to be of general use.
So merging the definitions leads to the following description:

Information Security is the science of assuring safety to a collection of data put into context.

1.2 Information Characteristics

Now that it is a bit clearer as what is the definition of information security I started wondering why information needs to be secured. In this section it is shown that information characteristics can be subdivided into three parts.
Ever since the development of the mainframe the C.I.A. triangle has been considered the industries' standard for describing the information characteristics that affect the value of information. The C.I.A. triangle describes three Critical Information Characteristics information possesses, videlicet confidentiality, integrity, and availability; hence the name. Definitions for the Information Characteristics as suggested by McCumber [2005] are adopted as they are to the point and reasonably clear.

Confidentiality

Confidentiality is the assurance that information access and disclosure is preserved for authorized persons. This includes the protection of personal privacy and proprietary information.
That confidentiality is an important information characteristic is clear because information can be seen as an advantage. And as a result, free information sharing will lead to a loss of advantage.
This fact was recognized early in history and the best-known example might well be cryptography that comes from the Greek words kryptós meaning hidden and gráphein meaning written. Cryptography already found wide use in the Roman Empire and its oldest form the substitution cipher is accredited to famous Roman general Julius Caesar. In substitution each symbol is replaced by another symbol resulting in an at first sight unreadable message. Which the reader then needs to decipher in order to be able to read it.
Modern ways to ensure that confidentiality is not breached are numerous, for instance some possible countermeasures are:

- The classification of information, viz. classified, secret, top secret.
- Information providence strictly on a need-to-know basis.
- Secure storage of classified data as for instance a clean desk policy.
- Utilization of general security policies.
- Education and awareness security programs for users.

Integrity

Integrity is the insurance that information is not altered or destructed in any improper way. Surely it is intuitively clear that integrity is a valuable information characteristic, and that if it is not maintained can certainly lead to loss of value for that information. As can be explained by the following statement: information can prove valuable, but disinformation can prove disastrous.

Possible ways to ensure integrity are authentication and nonrepudiation. These are loaded terms and are in this methodology considered as examples with which integrity can be established.

Authentication is a security control designed to establish the validity of a transmission, message, or originator or a mean of verifying an individual's authorization to receive specific categories of information. In general authentication can be verified in four ways:

- Do you know what you should know? This is maybe the best known form of authentication since one uses passwords for the withdrawal of money from ATM machines, alarms, computer profiles, etc.
- Do you have what you should have? This procedure is based on the fact that the user has possession of an artifact. Examples include credit cards, passports, smart cards and tokens.
- Do you have the features you should have? This area involves the use of biometrics such as fingerprints, retina scans, facial recognition and voice recognition. Biometrics is assumed to be one of the most effective forms of authentication.
- Do you do what you should do? This requires one to perform or produce something in an appropriate way. An example is signature recognition.

Procedures that require multiple authentications are considered *strong authentication*.

Nonrepudiation is the assurance that the sender of information is provided with proof of delivery and the recipient is provided with the proof of the sender's identity, so that neither can later legitimately deny having processed, stored or transmitted the information. This can be done with the use of asymmetric encryption or public key encryption that uses two encryption keys, one for encrypting and one for decrypting. Nonrepudiation can be established when the sender encrypts their information using the private key which anyone can decrypt by using the public key. In most cases the receiver in turn is asked for authentication. An example of its use is online banking in which case the receiver often is supposed to use some procedure of strong authentication.

Availability

Availability is the insurance that access to and the use of information is timely and reliable. This is maybe not such a straightforward information characteristic but consider the following example of option brokers. There are brokers whose job is to watch for the miss pricing of options on different markets due to non-overlapping bid-ask spreads. The first broker to find the gap will trade till the gap ceases to exist. Accordingly the sooner the information is available to a broker the sooner the trading can start and more money can be made.

To ensure that information is available some form of communication needs to be established. Internet is one of the fastest growing markets in communication and the supply of information. Other well established communication channels include the telephone/cellular and post.

1.3 How does Information Security affect me?

After reading the previous sections it is a bit clearer what information security means and why information is worth securing. Also, it was tried to suggest that there are examples in every day life in which most of us are confronted with security measures. But still the sharp-witted can make the analogy with the television that finds wide use and still the mechanical working of the machine is of no interest to most people.

Let me start of by saying the trivial that information security needs not affect one. If you live in a society where you do not need to bother with securing your information in any way reading this about this topic may be an entire waste of time. But if your everyday live or occupation involves the creation or handling of confidential information or you are in any other way interested in security it might be worth spending some effort.

I am convinced that information security and other branches of security in general are not so different at all. That is to say, I believe that all aspects of security such as communications, physical and personal security have a lot of similarities. Often safeguards contributing to one area of security also contribute to other aspects of security. Take for instance the following examples of security:

- The physical security of organizational properties. These include technical items such as hard disks, backups, servers, etc, but also physical objects as for instance confidential documents of areas such as the workspace.
- The personal security of employees that is designed to protect all the authorized personnel. Because one need not forget that often people are the weakest link. We all know too well the sinister tales of citizens abducted by secret agencies forced to give vital information.
- The security of communications. That focuses on keeping private company matters private irrespective of the medium or technology that is being used. One can think for instance of encryption of highly classified data such as customer credit card payments or the scrambling of telephone conversations.
- And information security. Which are information centered security measures such as nonrepudiation, authorization, and cryptology to name the most well known.

I think that these are convincing examples of the overlap of the different branches of security. This is not to say that some branches of security are superfluous, but is to stress that they are all different sides on the proverbial coin.
Anyways, to come back to the topic how information security can affect you personally the following examples are added:

- Who has not heard about the story of the PhD student who worked for two years on his dissertation only to find that after having his final draft ready the hard disk failed and the information on it could not be retrieved. How tragic though this story is the student forgot to consider the availability of his information. Surely relying on a single copy without backups is a disaster waiting to happen!
- The famous PIN code written on the back of your check card. Only to find that you have lost your wallet and that someone has cleaned out your saving account. Where in this case the confidentiality perspective was forgotten.

- Or the person who gave his/her EBay account and number away to a spoofed mail, and which as a result got the bill for all kinds of items he did not order. Here the integrity aspect of the sender was not considered, and, the mistake was made to hand out confidential information.
- The password of you computer written on a paper taped under the keyboard, which can cause serious confidentiality problems.
- Opening an email attachment and afterwards noticing that the attachment acted as a Trojan horse and that your computer now is infected with viruses. This is yet another striking example of a problem where the integrity of the information was not confirmed and resulted in an availability failure.
- Turning off the automatic update of your antivirus resulting in the infection of your computer by the newest virus. This example also seriously affects the availability of your information.

As you may have already noticed information security is for a large part a result of the information age in which the movement of information became faster than physical movement and started roughly post 1970.

1.4 Information States

In the previous sections it was shown what information security is, that information has an intrinsic value and examples were given why it should apply to you. But it is one thing to know that you need security, it is of no use at all if you do not know what to secure.
This may well be considered as one of the most difficult parts of information security: information is not really tangible since information can take so many different shapes. And if the information is already so impalpable the security of that information adds another layer to the complexity.

A few well known examples of *information states* are the following: The storage of information which can be decomposed as short-term storage such as RAM (Random-Access Memory) or long term storage such as hard disks or DVD's (Digital Versatile Disk), the processing of information by a processor, the transmission of information over the ether or broadband and the display of information on your laptop.

For purpose of generality, the following three information states are adopted: storage, processing, and transmission. Certainly the problem accompanying displaying information is a hot topic and is by definition not easy to protect against. To my opinion shoulder surfing is not professional but it will be a cold day in hell when we will see a prosecution for the act. My argument to adopt only the three information states is that the vast majority of security breaches occur as breaches in storage, processing or transmission.

To indicate how valuable these information states are just think of the different ways of securing them. You can encrypt information stored on a hard disk but it is impossible to process or display encrypted information without decrypting it first. The problem of shoulder surfing could be contained by adopting policies for restriction access but one surely cannot restrict transmission as that is perpendicular to the purpose.

1.5 Who poses a threats?

Another question that is appropriate to ask is who to protect information from. It may be clear that the intentions for attacks are quite different in nature. Some are circumstantial and may not even be intended against the target specifically, but others may be directed. Circumstantial attackers can often be discouraged easily with minor precautions, whereas directed attacks may be very expensive to prevent and in many cases nearly impossible.

One difference between attackers often lies in the intentions. Nichols, Ryan and Ryan [2000] for instance differentiates between information attacks which occur because of negligence or an accident as for instance the act of God, random attack by for instance vandals, attacks for prestige of the success by hackers, attacks for personal gains by outsiders or insiders, attacks for strategic advantages by competitors, attacks against the society you happen to live in by terrorists or armed forces, those who attack you to get to someone else, and so on.

An insider attack is often more severe than an attack of hackers since hackers often do not have the motivation to hurt the company but for the prestige that comes from the successful attack itself. An inside attack on the other hand is done by a person who is trusted and has additional information about the organization, and as a result can do significantly more harm.

Attacks by competitors are often also very difficult to protect against since attackers could be both insiders and outsiders. Furthermore, these attacks are likely to involve well-trained attackers since the stakes will typically be high for the attack to occur in the first place. Well known examples include money laundering and industrial espionage. But also, in case of direct competition, attacks could be directed at slowing down your decision cycle. Such an attack would be difficult, if not impossible to detect.

Post 9/11 terrorist attacks are receiving a lot of attention. But, regardless of all the attention it remains hard for companies to assess if they are targets. It is so hard because one needs to think from the attackers' point of view. One rule of thumb is that your company might well be a target if your company is perceived by the terrorists as in support of the government or does an attack on your company in other ways contribute to the status of the terrorists.

1.6 Controls and countermeasures

The most important question still remains to be answered: is there not something that can be done to counter the attacks in the first place? Fortunately there are, and in the information security realm they are labeled as controls, countermeasures, or safeguards.
Roughly, countermeasures can be divided into three groups:

- Technology based safeguards such as firewalls, cryptology and IDS.
- Management based safeguards that are often based on policies and procedures. Examples include hiring and firing policies and checking in and checking out of company premises.
- Mitigation based safeguards as the engineering/architecture of floor plans, insurance of goods and training of awareness among employees to name but a few.

The difficulty with implementing safeguards is that you preferably would try to protect against every single threat. And to secure against a threat one first has to acknowledge the threat, such that again information may well be your most valuable asset.
The key in securing is that you basically want to secure the entire spectrum against threats. Take for instance the spectrum obtained by looking at information characteristics and information states. Heavily fortifying your position of confidentiality in the storage state is needles to say very prudent. However, if, as a result no money is available to secure information confidentiality in processing state than the strong security in the one direction may well be in vain since the determined thief will probably take the open backdoor.
There is the analogy with the chain which fails due to the weakest link, and, as is known there only needs to be one.

Another key principle in information security, is the term “layered security”. An instructive example might well be that you want your information protected from every possible threat be it a circumstantial attack, an insider attack or an attack by a hacker. Since they are all different in nature, the countermeasures that could be taken against each attack are different.
A notional indication of layered security and how it could work in this example is given in Figure 1. In this figure the solid lines represent the different layers of security and the dotted lines represent different kind of attacks. Note how the different layers of security are safeguarding against different kind of threats.
The notional picture of layered security is not new, but the triangles are often replaced by a set of circles. To my opinion that is a bit confusing since every side on the circle looks the same, whereas a triangle for instance has three sides that can be specifically pinpointed. Moreover, different sides could be identified with for example different information characteristics.
Ideally, in this example multiple 9-sided surfaces could have been chosen - one for every information characteristic and information state combination - but unfortunately my drawing skills ruled out this possibility. Since the purpose of this picture was to make the point it was chosen not to take such drastic steps as to get familiar with an advanced drawing program.

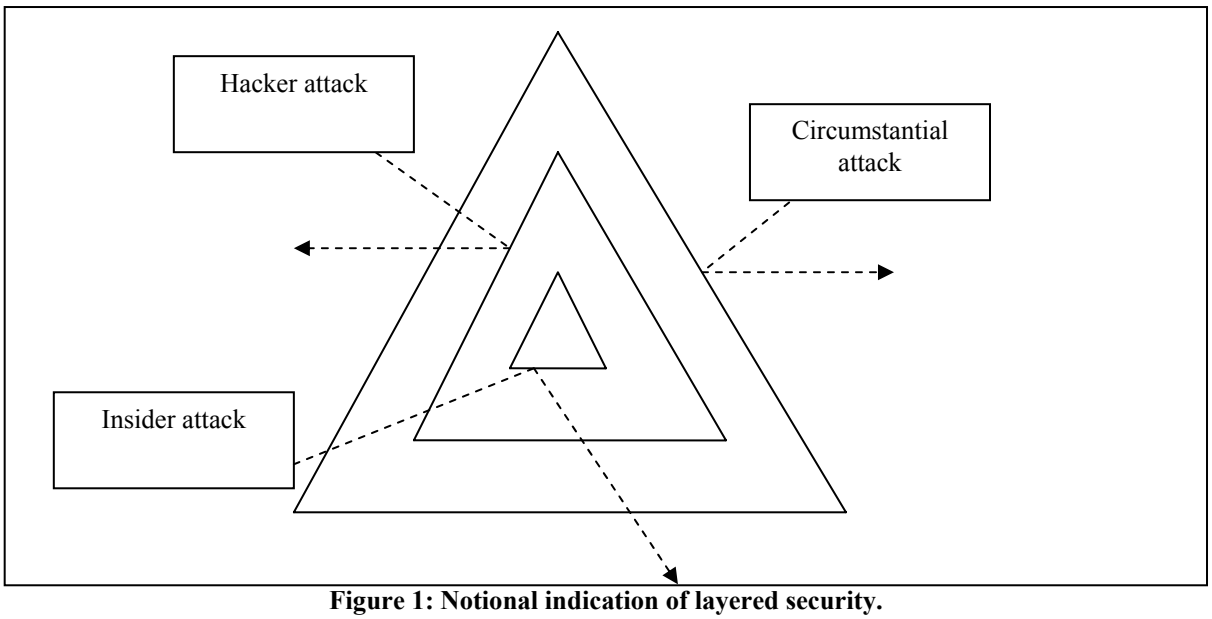

Figure 1: Notional indication of layered security.

1.7 Phasing security

Obviously to decrease the chances of being attacked some level of protection is worthwhile. But since the protection cannot guarantee safety other precautions have to be taken. A familiar phrase in information security is protect, detect and correct. The protection, detection and correct phases are known to be the three security phases.

The protection phase is regarded as the most basic phase in information security. Who doesn't have a firewall or antivirus running on their computer? Firewalls, antviruses and proxy servers are in information security often basic precautions for individual systems that deter circumstantial attacks by hackers but in general they are not able to guarantee full security.

Since there is always the possibility that an attack has occurred the detection phase is of great significance. The ability to detect if an attack has happened may be of vital importance to show that the system is working properly. Technical precautions that can be taken involve different kind of intrusion detection systems (IDS). These IDS's are more elaborate and involve separate systems that monitor the network such as network-based IDS or computers such as host-based IDS.

Without the correction phase, however, detecting that a violation has occurred is pointless. This phase may include technical precautions such as the existence of backup storage facilities and hot or cold sites. But also procedures to get the facilities online as soon as possible is an important countermeasure. Furthermore, the phase might require procedures to deal with attacks while in progress.

2 Overview of the Poisson Process a Stochastic Process

The model that is used in chapter 3 to describe attack patterns from cyber attacks is based on the Poisson process. In this chapter the formal mathematical underpinning for the Poisson process are given.

Section 2.1 elaborates on stochastic processes of which the Poisson process is but an example. Then in section 2.2 the relation between counting processes and Poisson processes are treated and definitions are given. In section 2.3 homogeneous Poisson processes is treated, and last but not least the non-homogeneous Poisson processes is discussed in section 2.4.
For readability purposes the proofs of the various theorems and statements have been placed in the appendix at the end of the thesis.

2.1 Stochastic processes

The non-homogeneous Poisson process (NHPP) is included in the realm of stochastic processes. According to mathematical definitions a stochastic process is an indexed set of random variables defined on the same probability space taking values in the same codomain. You can just think of them as random functions.
Domains over which stochastic processes are defined include time and space. In these applications the stochastic processes are often referred to as time series or random fields, respectively.
Popular examples of the practical use of time series analysis include stock market and exchange rate fluctuations. These are often modeled as Brownian motions that are usually thought of as random walks.
In Figure 2 the price development of a share of Royal Dutch Shell is shown. Option traders price options on stock based on the assumption that the stock can be modeled as a Brownian motion. The price development over past years in the terminology of stochastic processes is considered to be a sample path.
Another example an application of stochastic processes is martingale theory in which the processes are constrained on the expectation. To be more precise a stochastic process is a martingale if the conditional expectation of a random variable given the past observations equals the last observation.
Another example in which stochastic process are used is Markov theory in which the domain and codomain are labeled state spaces and where the conditional probability of the next state space can be considered as a random function of the last state space alone.
And last but not least a Poisson process is an example of a counting process that is yet another subset of the stochastic processes. It is hoped to be able to show that the Poisson process can, in addition to numerous other applications, be used to model cyber attacks and the engineering of the protection from them.

Figure 2: Development of a share of the Royal Dutch Shell on March 30 2005 over the past year, in €. Source: http://www.belegger.nl/

2.2 Counting processes

Let us consider the non-negative random variables $T_1, T_2, \ldots$, for $T_i \in [0,\infty)$ with respective arrival times $t_1, t_2, \ldots$. Then also the number of arrivals up to time t is a random variable. A point process or counting process, usually denoted $N(t)$, is a stochastic process which counts the number of arrivals in an interval $[t_0,t)$ and is thus a random variable. An example of such a process is given in Figure 3. Note that the counting process is left continuous.

Without loss of generality it is assumed that $t_0 = 0$ and $N(0) = 0$ and that the sequence of arrival times is ordered, that is to say $t_1 < t_2 < t_3 \ldots$ Then a counting process is defined as follows.

Definition 2.2.1 A *counting process* (point process), $N(t)$, is defined as

$$N(t) = \max_n \{n : t_n \leq t\} \tag{2.1}$$

The homogeneous Poisson process and non-homogeneous Poisson process are the most commonly used counting processes in reliability. And since Poisson processes are point processes they count the number of events of a certain type in a specific time interval. The NHPP is a generalization of the HPP, but the HPP is intuitively clearer and therefore it was chosen to start off by explaining this process.

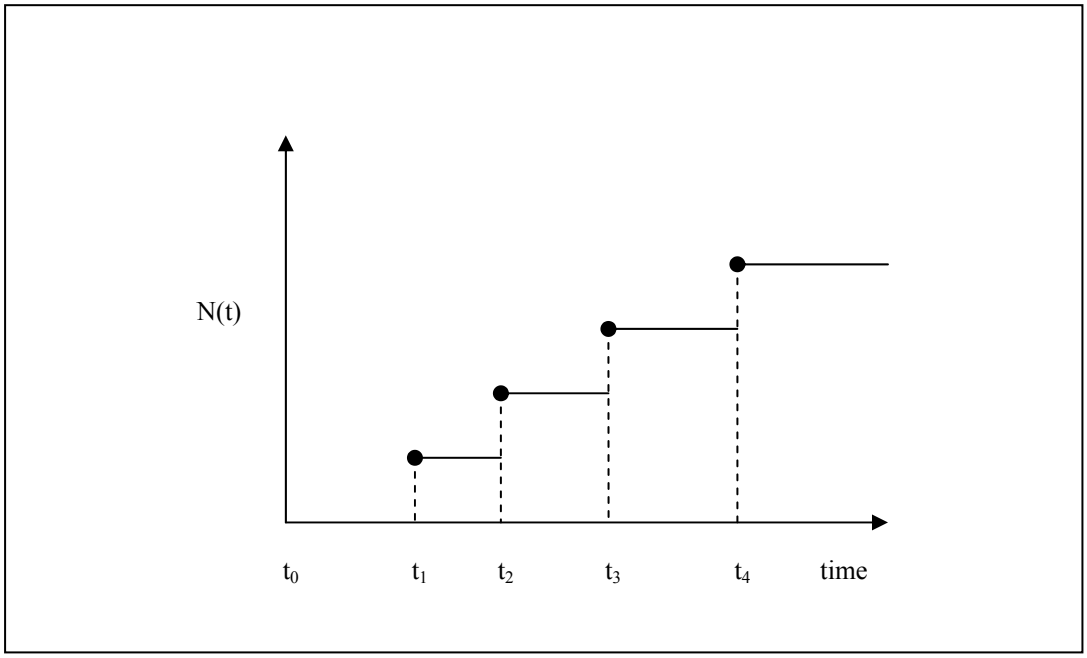

Figure 3: Counting process *N(t)*.

2.3 The homogeneous Poisson process

The HPP is also one of the most frequently used stochastic processes. It owes its name to the famous French mathematician, geometer and physician Siméon-Denis Poisson (1781-1840). Other renowned works include the correction of Laplace's partial differential equation the Poisson equation, series of memoirs on definite integrals, and discussion of Fourier series.

The HPP is an example of a counting process. The distribution of the process is well known and less stringent ways to derive the probability density function exist and of which the following is just an example.
An interesting way to derive the distribution of the HPP process is via the binomial distribution that describes the number of successes of n independent yes/no experiments and taking the limit for n to infinity.
Thus consider the variable X that follows a binomial distribution with parameters n and $\lambda t/n$ then according to the law of rare events

$$\lim_{n\to\infty} Pr\left(X=k\right) = \lim_{n\to\infty}\binom{n}{k} p^k \left(1-p\right)^{n-k} = \frac{\left(\lambda t\right)^k}{k!} e^{-\lambda t}. \tag{2.2}$$

A proof of which is supplied in the appendix.

Of course the probability distribution can also be obtained from the more formal definition of the HPP that is given below.

Definition 2.3.1 A counting process $\{N(t): t \geq 0\}$ is called a homogeneous Poisson process if the following conditions hold

1. $N(0) = 0$,
2. The increments of the process are stationary and moreover the increments of disjunct intervals are independent,
3. $P(N(t+\Delta t)-N(t) = 1) = \lambda \Delta t + o(\Delta t)$,
4. $P(N(t+\Delta t)-N(t) \geq 2) = o(\Delta t)$.

Where it is noted that $\lambda > 0$ and constant. The parameter lambda is often referred to as the rate, or intensity, or rate of occurrence of failure (ROCOF) of the process. Furthermore $o(\Delta t)$ denotes that $o(\Delta t) / \Delta t \to 0$ as $\Delta t \to 0$.

The definition above may seem a bit awkward to any non-mathematician since on first hand it seems that the probability distribution is not specified. Fortunately theorem 2.3.2 states that the distribution function is specified.

Theorem 2.3.2 If $\{N(t): t \geq 0\}$ is a HPP with parameter $\lambda > 0$, then the probability of observing k arrivals in the interval $(0,t]$ is Poisson distributed and given by

$$P\left[N\left(t\right)=k\middle|\lambda\right] = \frac{\left(\lambda t\right)^k}{k!} e^{-\lambda t} \qquad k = 0,1,\ldots \tag{2.3}$$

The proof of theorem 2.3.2 is a bit more extensive compared to the proof of Equation (2.2), and, moreover, is more mathematical in nature. The more interested reader is referred to the appendix where a proof using generating functions is supplied.
Note that the result also works in the other direction. If the probability density of a process of observing *k* arrivals in the interval *(0,t]* is Poisson distributed, then there exists a HPP governing the process.

For the layman in the field of mathematics and HPP's corollary 2.3.3 is supplemented stating both expectation and variance of a Poisson distribution.

Corollary 2.3.3 The expectation and variance of a Poisson distribution with parameter λ are given by

$$E\left[N(t)\right] = \lambda t$$
$$Var\left[N(t)\right] = \lambda t. \tag{2.4}$$

The proof of corollary 2.3.3 using the generating function representation is supplemented in the appendix.

A main reason why the HPP is one of the most used counting processes may well be the property that the superposition of two independent homogeneous Poisson processes is yet another homogeneous Poisson process. An illustrative example of such a super-positioned Poisson process is given in Figure 4.
In fact the relations are even stronger, it turns out that the rate of the superimposed process is the sum of the rates of the two processes as is stated in theorem 2.3.4.

Theorem 2.3.4 Consider two independent homogeneous Poisson processes, $N_1(t)$ and $N_2(t)$, with intensities λ_1 and λ_2, respectively. Then the superimposed process is again a homogeneous Poisson process with intensity $\lambda_1 + \lambda_2$.

The proof is straightforward and is given as a supplement in the appendix. Note that a general result from theorem 2.3.4 is that the superimposition of *n* independent HPP's results in a HPP with as intensity the sum of the *n* independent HPP's.

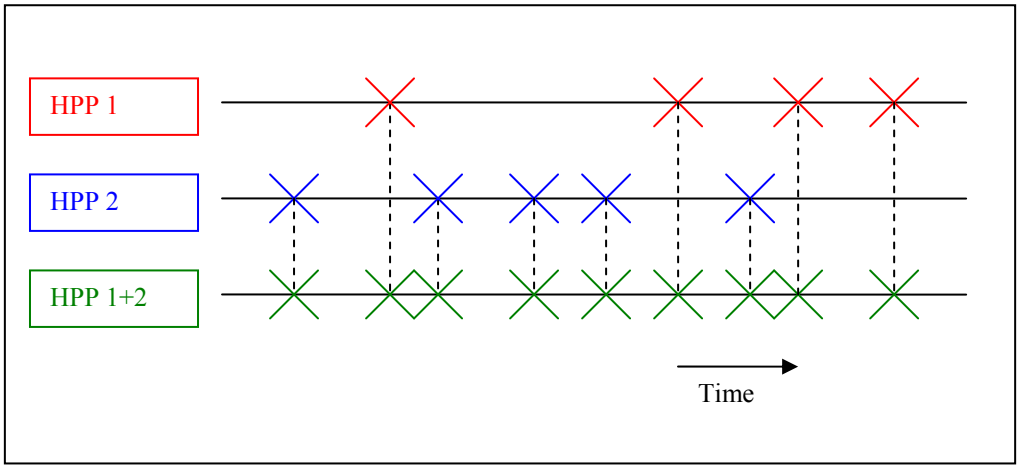

Figure 4: Superposition of two HPP's leading to a new HPP.

A question that might have come to mind after interpreting Figure 4 could well have been: what does a typical arrival pattern of a HPP look like? Corollary 2.3.3 states that in case of a HPP with rate λ the expected number of failures in any interval $[s,t)$ where $s < t$ only depends on the length of the interval and is simply $(t - s)\lambda$. It may however still remain unclear what a typical pattern of arrival times look like. Do the arrival times cluster around a specific point or are they equally distributed over the range?
Luckily, conditional on the number of failures, it can be proven that the arrival times are uniformly distributed over the support as is stated in theorem 2.3.5.

Theorem 2.3.5 For $\{N(t): t \geq 0\}$ a HPP and let $N_T = n$ on the interval $[0,T)$, then conditioned on the number of arrivals the n arrival times are independent uniformly distributed on the support.

Once again interested readers are referred to the appendix for the proof.

From theorem 2.3.5 it follows that if the conditional distributions are expected to be evenly distributed over the entire range a general pattern of a HPP over time interval $[0,T)$ in general is equally distributed over the entire range.
Let's consider again the superimposed HPP and assume that the total number of failures is known but not the number per independent Poisson process, and, furthermore assume that the intensities of the superimposed processes are known. Then the probability distribution of the number of failures per Process is multinomial distributed as corollary 2.3.6 states.

Corollary 2.3.6 Consider two independent HPP's with $N_1(t)+N_2(t) = m$ and the intensities of the HPP's are λ_1 and λ_2, respectively. Then the probability that $N_1(t) = n$, $n = 0,1,\ldots,m$ is given by

$$P\left(N_1(t) = n \middle| N_1(t) + N_2(t) = m\right) = \binom{m}{n}\left(\frac{\lambda_1 t}{\lambda_1 t + \lambda_2 t}\right)^n \left(\frac{\lambda_2 t}{\lambda_1 t + \lambda_2 t}\right)^{m-n} \tag{2.5}$$

The proof consists of applying the definition of the conditional distribution and using independence of the processes. For completeness it is given in the appendix. Note that it follows that for a superimposed process of say k independent processes a k-dimensional multinomial distribution is obtained.

A lot of properties of the homogeneous Poisson process have been treated in this part. In the next section it is shown that the non-homogeneous Poisson process, where the assumption of a constant intensity is dropped, shares a lot of the same properties.

2.4 The non-homogeneous Poisson process

A non-homogeneous or inhomogeneous Poisson process is a Poisson process with a non-constant intensity. Actually the HPP can be thought of, as a special case of the NHPP but the choice in this paper has been to consider the NHPP as a generalization of the HPP.
In this section it will be shown that there exist NHPP counterparts for many of the properties of the HPP shown in section 2.3. The main difference being that the accumulated intensity in the inhomogeneous case has an integral representation, which in the homogeneous case

simplifies into a multiplication of period and intensity. This renders somewhat disagreeable formulas.

The definition for a NHPP is basically the same as the HPP with the adjustment that the intensity is a function of time.

Definition 2.4.1 A counting process is called a non-homogeneous Poisson process with arrival rate $\lambda(t) > 0$, if

1 $N(0) = 0$,
2 Non-overlapping increments are independent,
3 $P(N(t+\Delta t)-N(t) = 1) = \lambda(t)\Delta t + o(\Delta t)$,
4 $P(N(t+\Delta t)-N(t) \geq 2) = o(\Delta t)$.

Note that as a result of the non-constant arrival rate, the stationarity of the increments is lost.

The probability distribution of the NHPP is fixed by definition 2.4.1 and is known to follow a Poisson distribution. Theorem 2.4.2 states the precise parameters of the model.

Theorem 2.4.2 If $\{N(t): t \geq 0\}$ is a NHPP with parameter $\lambda(t) > 0$, then the probability of observing k arrivals in the interval $[0,t)$ is Poisson distributed and given by

$$P\left[N(t) = k \middle| \lambda(t)\right] = \frac{\left(\int_0^t \lambda(\tau)\,d\tau\right)^k}{k!} e^{-\int_0^t \lambda(\tau)d\tau} \qquad k = 0,1,\ldots \tag{2.6}$$

The proof is supplemented in the appendix and is basically the same as the proof of theorem 2.3.2. Note how the intensity integration becomes a multiplication of time and intensity in the case of constant arrival rates.

One difference with the homogeneous Poisson process is that the distribution for the number of arrivals in an interval $[s,t)$ with $t > s \geq 0$ does not have to be equal to the number of failures in the interval $[0,t - s)$ because the stationarity assumption was dropped. The distribution function can be easily calculated with the aid of the moment generating function and is given in Theorem 2.4.3.

Theorem 2.4.3 Let $\{N(t): t \geq 0\}$ be a NHPP with parameter $\lambda(t) > 0$ and $t > s \geq 0$, then the probability of observing k arrivals in the interval $[s,t)$ is Poisson distributed and given by

$$P\left[N(t) - N(s) = k \middle| \lambda(t)\right] = \frac{\left(\int_s^t \lambda(\tau)\,d\tau\right)^k}{k!} e^{-\int_s^t \lambda(\tau)d\tau} \qquad k = 0,1,\ldots \tag{2.7}$$

The proof is completed by noticing that the moment generating function of Equation (2.7) is the fraction of the individual moment generating functions, see appendix.

The expectation and the variance of the NHPP directly follow from corollary 2.3.3 and are both given by the parameter of the Poisson distribution. Note that in case of the homogeneous Poisson process the expectation and variance are not stationary, this is a result of the fact that the process itself is not stationary.

A property that is preserved for the inhomogeneous process is the superposition property. The following theorem states that the superposition of two independent NHPP's $N_1(t)$ and $N_2(t)$ is also a NHPP and the rate of the superimposed process is the sum of the rates of the processes $N_1(t)$ and $N_2(t)$.

Theorem 2.4.4 Let $N_1(t)$ and $N_2(t)$ be two independent NHPP's with arrival rates $\lambda_1(t)$ and $\lambda_2(t)$, then the superimposed process is again a NHPP's with arrival rate $\lambda_1(t) + \lambda_2(t)$.

The proof follows from conditioning on the events and rewriting terms. Note that from theorem 2.4.4 it follows that the sum of n independent NHPP's itself is a NHPP as arrival rate the sum of the n arrival rates and is supplied in the appendix.

A difference between the two processes is the conditional distribution. Where the HPP conditional distribution was uniform over the support the NHPP is not. Its conditional distribution is the hazard rate scaled to one as is asserted in theorem 2.4.5.

Theorem 2.4.5 For a NHPP's $N(t)$ with intensity $\lambda(t)$ conditioned on $N(T) = 1$, the conditional distribution on $[0,T)$ has the following density function

$$f_{t_1}(t) = \frac{\lambda(t)}{\int_0^T \lambda(\tau)\, d\tau} \tag{2.8}$$

The proof can be found in the appendix

Consider again the independent superimposed NHPP and assume that the number of failures of the superimposed process is known, and, furthermore the individual arrival rates are known. Then the distribution of the number of failures per process is multinomial distributed as is shown in the next corollary.

Corollary 2.4.6 Let us consider two independent NHPP's with $N_1(t)+N_2(t) = m$ with intensities $\lambda_1(t)$ and $\lambda_2(t)$, respectively. Then the probability that $N_1(t) = n$, $n = 0,1,\ldots,m$ is given by

$$P\left(N_1(t) = n \middle| N_1(t) + N_2(t) = m\right) = \binom{m}{n}\left(\frac{M_1(t)}{M_1(t)+M_2(t)}\right)^n \left(\frac{M_2(t)}{M_1(t)+M_2(t)}\right)^{m-n} \tag{2.9}$$

where

$$M_1(t) = \int_0^t \lambda_1(\tau) d\tau,$$
$$M_2(t) = \int_0^t \lambda_2(\tau) d\tau. \tag{2.10}$$

The proof can be obtained by replacing $\lambda_1(t)$ and $\lambda_2(t)$ in the proof of corollary 2.3.6 with $M_1(t)$ and $M_2(t)$, respectively. Note that it follows that the conditional distribution of the number of failures per distribution resulting from k independent, super positioned NHPP's follow a k dimensional multinomial distribution.

3 Model for Cyber Attacks

The concept of cyber attacks comes from the information warfare real, see for instance Nichols, Ryan and Ryan [2000]. It is very general and based on the notion that information is the basis for calculated decision making. As such a cyber attack is in general designed to disrupt or delay the decision making process of rivals, this can be done by a direct attack on information systems, an indirect attack on personnel maintaining the facilities or on the decision making process itself.

For the rest of this chapter it was chosen to focus on the theoretical underpinning of cyber attacks taking place in the information security domain. I have modeled the attacks to get numerical results in a domain in which quantitative results are scarce. It is hoped that the results underwrite intuition and experience from experts in the field of information security.

It is proposed that the number of successful attacks from a particular type at any time t is Poisson distributed. By adjusting the arrival rates it was tried to model different kinds of attacks in the area of information security. Furthermore, it is proposed to use a log-linear model to incorporate the use of different kinds of safeguards

3.1 A virus as an attack

The choices of possible attacks on information systems are enormous. There are physical attacks that can be directed to the information systems themselves, examples of those include bombs and theft but may even include abductions and acts of god. Furthermore, there can be non-physical attack such as Trojan horses, spies and viruses.

In this part the example of viruses is examined further since networks have been recognized as the battlefield of the future, see for instance Ryan, Woloschek, and Leven [1996]. In subsection 3.1.1 an overview of models for virus propagation and virus detection is given. In subsection 3.1.2 the model for the viruses themselves is developed and an example is given in which the model is tested. And finally in subsection 3.1.3 a model for safeguards is introduced and an example shows how different patch-times significantly affect the number of successful attacks.

3.1.1 Propagation of viruses

It has been suggested that the study of the spread of malicious code will enable one to make the defense against that spread more resilient. Recent studies examine the propagation of viruses and their relation to information security.
A lot of epidemiological based studies use either the Susceptible-Infected-Susceptible (SIS) model or Susceptible-Infected-Removed (SIR) model. Such models determine virus spread over time and may indicate thresholds for viruses to result in epidemics. The SIS model only supposes that a system from a susceptible state may enter an infected rate and vice versa. Whereas the SIR model supposes that systems can enter the removed state either because the

system has been removed because of failure or because the system has become immune after recovery.
The drawback from these models is that systems immediately enter the infected state from the susceptible state. In reality, however, there may well be delays because of user inactivity, computer configuration settings or it might even be that it is a property implemented in the virus because of stealth considerations.

Thommes and Coates [2005] have modeled the propagation of infected files in a Peer-to-Peer network by examining differential equations for hosts. The model is based on the S-E-I-R model used in epidemiology to study the spread of diseases in a population. All the individuals in the population are classified as either Susceptible, Exposed, Infected or Recovered, so that hosts do not immediately have to infect others but may reside in the exposed state for some period of time. In addition, the transitions of the states of the individuals are described by differential equations. With this model state evolutions over time were obtained both for the closed form solutions of the differential equations and their numerical counterparts calculated with Euler's method. Furthermore, the model was able to model the effect of online/offline behavior of users.

Wang and Wang [2005] also considered epidemiological models and found that timing parameters as infection delay and user vigilance may influence propagation. An infection delay is the delay from the time that a system is infected to the time that the system acts as a host itself and user vigilance is the time period after recovery from infected state that the system stays immune before returning to the susceptible state. It was found that infection delay and user vigilance can contribute to the ability to contain viruses.

But what is seen in the epidemiologic models is that all the susceptible systems can become infected given the right conditions. Much speculation exists if the monopoly of Microsoft enhances the risk of a critical system failure. In Ryan [2005] it was found that when 43% of a network is susceptible to attacks, a single virus could cause a critical attack, and if the network was highly connected, the susceptible threshold was a mere 17%. Fortunately the dominance of Microsoft is only in the reigns of operating systems, by as much as 97%, and not on the Internet as a hole. This research did however indicate the fragile nature of monocultures in information infrastructure.

The epidemiological approach however is only one of many as I hope to show. One of the downsides of the epidemiological approach is that systems are at the start either labeled susceptible or infected, and as a result systems can only be infected once. This doesn't seem to agree with reality as viruses could infect systems on multiple occasions.

Other approaches focus on the combat against viruses on a single system or small network scale.
At the present time virus detection systems are commonly based on pattern behavior of data. Some detection systems focus on the patterns in the header information of incoming packages or on programs that dwell in the computer, commonly referred to as network-based intrusion detection systems and host-based intrusion detection systems. The patterns of behavior of the system are then compared to signatures from known viruses.

In Ryan [2005] it is said that the Slammer Worm infected 90% of the vulnerable hosts within 10 minutes and is a typical proof of how inadequate the current virus detection systems are.

New virus detection systems that are not based on signatures may well be needed to counter the growing threat of virus infection.

Goel and Bush [2005] have compared virus propagation of computer viruses to the propagation of biological viruses. Several examples of comparisons between biological and information security concepts are described such as a defensive code that act as antibodies and turn off malicious code and dividing the network into pathways as in protein pathway mapping to be able to determine the root cause of events.
The authors used the biological paradigm of the immune system to transcend the specific signatures based virus detection systems. To cover the wide range of possible infections the immune system uses, among others, detectors with low specificity, which have the obvious downside that there is poor discrimination ability in the indicators.

Wehner used such a low specific indicator when analyzing worms in network traffic using compression. First, a Network Compression Distance (NCD) was defined that can be intuitively interpreted by the magnitude with which data can be further compressed. Then the ratio with which the data could be compressed further was calculated. This ratio was then the indicator for making the distinction between normal code and malicious code.
Distinctions between different networks protocols had to be made because of the difference in compress ratio. SSH sessions, for instance, are hardly compressible since they already carry encrypted payloads.
Means and variance were calculated for the different protocols and bounds were computed which covered the ratio's of the legitimate commands. The findings were implemented into the open-source IDS Snort [2005].
In the tests cases the program was able to detect new viruses with the low specific indicator. However, discrepancies between legitimate programs and malicious code could not always be made.

3.1.2 Model for viruses

In this section it was chosen not to assume a particular form for the propagation although it is interesting to see that the propagation models have some properties in common with the choice of the particular intensity studied in this model.
In this section it is assumed that a particular system is susceptible to the virus that is being spread by other infected systems, see Figure 5. Note how systems vulnerabilities depend upon the number of hosts for the virus but that the exact state of the other systems are not assessed such that the problem of defining states was avoided.
Furthermore, it is assumed that the number of successful attacks is Poisson distributed for which a time varying intensity is chosen. It is assumed that the arrival rates have a specific general linear form that was selected by considering the properties of viruses in general. The advantage of this model is that multiple attacks by viruses can occur as opposed to the epidemiological models.

Assume that at time $t_0 = 0$ there is no hazard but that at different stages new viruses are developed. It might be argued that t_0 in reality does not exist but it can be counter argued that it was somewhere before the development of the computer. In which case the probability that your computer is attacked would have certainly been zero, by the simple fact that one can not attack what is not there.

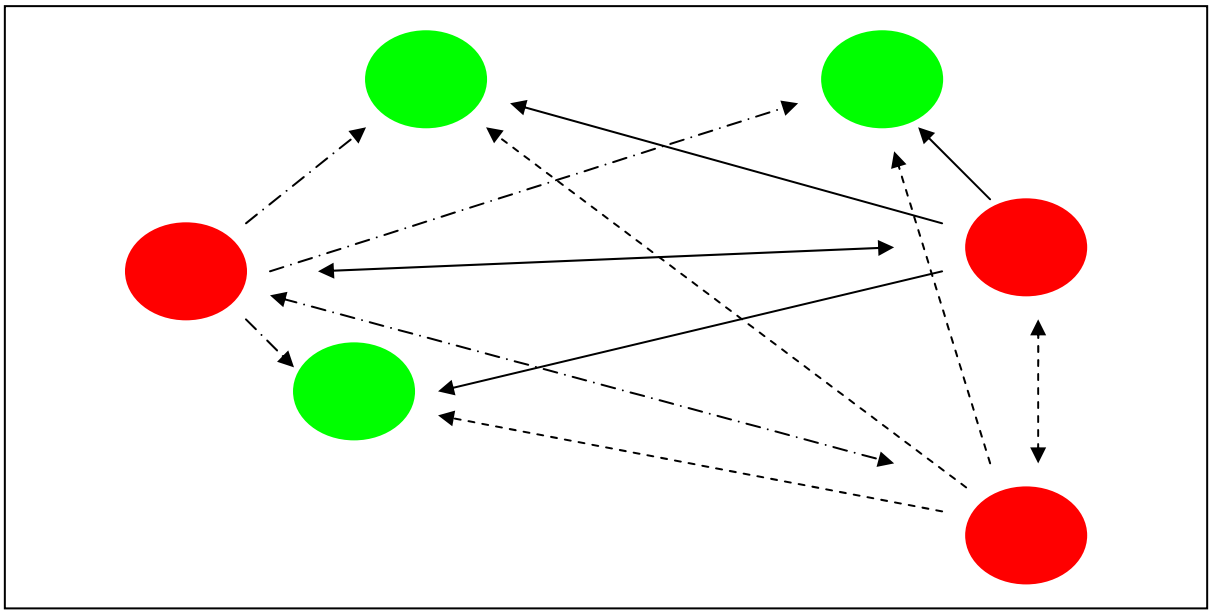

Figure 5: General propagation model for viruses. Red systems are infected and green systems are uninfected. Arrows indicate possible infection paths the virus could take to infect other systems.

Now look at the arrival rate of the *i*-th virus that is initiated at t_{i1}. Assume that the arrival rate of attacks is linear and is defined by

$$\begin{aligned}
&\tilde{\lambda}_i(t) = 0, && \text{for } t < t_{i1} \\
&\tilde{\lambda}_i(t) = m_1\left(\frac{t - t_{i1}}{t_{i2} - t_{i1}}\right), && \text{for } t_{i2} > t \geq t_{i1} \\
&\tilde{\lambda}_i(t) = m_1, && \text{for } t_{i3} > t \geq t_{i2} \\
&\tilde{\lambda}_i(t) = m_1 - (m_1 - m_0)\left(\frac{t - t_{i3}}{t_{i4} - t_{i3}}\right) && \text{for } t_{i4} > t \geq t_{i3} \\
&\tilde{\lambda}_i(t) = m_0, && \text{for } t \geq t_{i4}
\end{aligned} \tag{3.1}$$

where t_{i1}, t_{i2}, t_{i3}, t_{i4}, m_1 and m_0 are all constant such that $t_{i1} \leq t_{i2} \leq t_{i3} \leq t_{i4}$, $m_1 > m_0 > 0$. The different times can be interpreted as follows. Time t_{i1} is the initial time when virus *i* is released in cyberspace. Between time t_{i1} and t_{i2} more systems are infected and start acting as hosts for the virus which leads to an increase in intensity. Then at time t_{i2} the number of infected systems – and thus the number of hosts – is saturated at the highest level m_1, that is to say the rate at which the systems are getting infected is equal to the rate at which systems are being restored to their status before infection. At time t_{i3} the hazard of virus type *i* is decreasing, which may be a result of patches the software companies have come up with, change of operating environments or changes in technologies since all these changes result in the decrease of infected systems. Finally at time t_{i4} the use of the virus has dropped to the lowest level m_0, at which time some systems still are infected thus providing a relatively small hazard. The general form of the hazard is depicted in Figure 6.

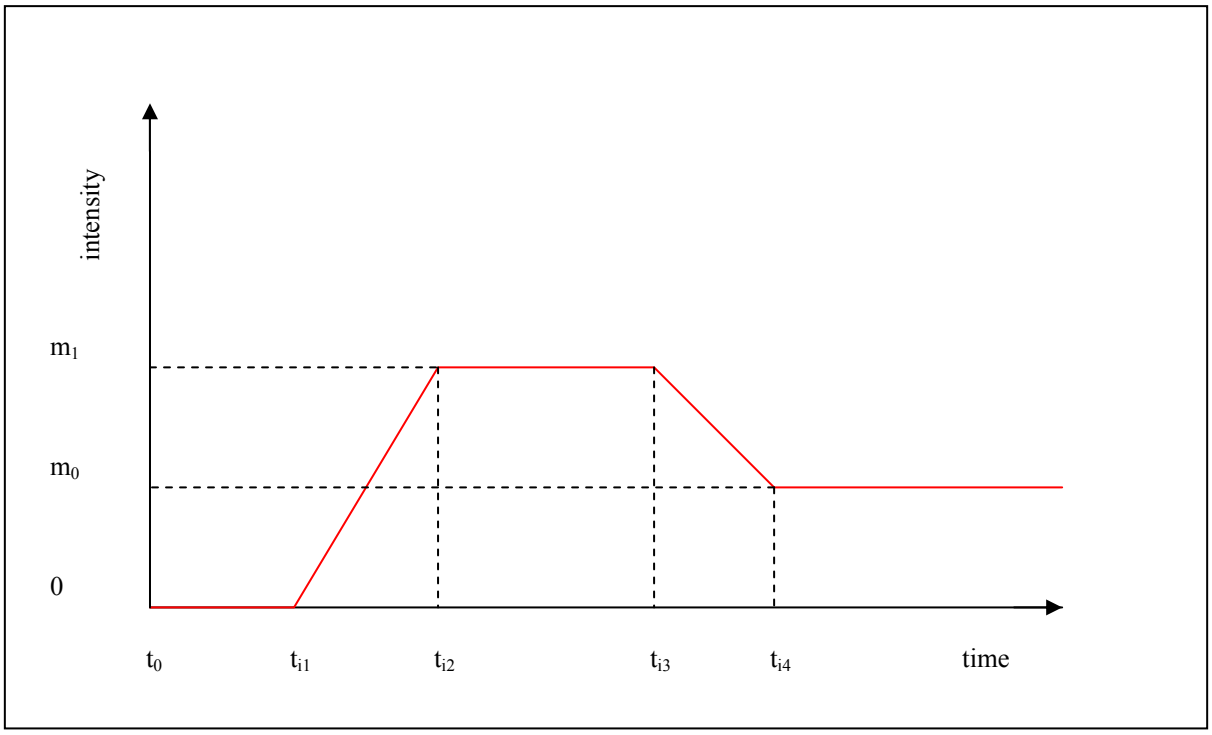

Figure 6: Illustration of linear intensity defined by equation (3.1).

It was observed by experts that the time between t_{i2} and t_{i3} in reality might be slim or even nonexistent. However, it is noted that time t_{i1} and t_{i2} in this model still may be chosen very close to one another and from a theoretical point of view it makes sense to distinguish between time of saturation and time of development of the patch.

3.1.3 Safeguards for viruses

In this example it is shown how the initiation of a new virus under the model assumptions influences a system that has no safeguards. The parameters to calculate the arrival rate of the successful attacks have been given in Table 1. To make the necessary calculations Excel was used and spreadsheets were computed.

Table 1: The data for virus 1.

Virus 1	
t_1	day 5
t_2	day 15
t_3	day 25
t_4	day 40
m_1	0.1
m_0	0.01

From Table 1 it is seen that at day 5 the virus is initiated and that at day 15 the number of the hosts for the virus is saturated. Then on day 25 the number of hosts for the virus decreases until day 40 when the intensity stays constant until the end of the study at day 150.

Since pictures are supposed to say more than a thousand words I have supplemented a plot for the intensity and for the expected number of successful attacks in Figure 7 and Figure 8, respectively.
Note that the expected number of successful attacks grow rapidly during the first 40 days when the intensity is relatively high. After day 40 when the attack type is already 'old' the expected number of failures grows linearly and only slightly.

Another interesting figure is the probability distributions for the number of successful attacks over time as is shown in Figure 9. In which the probability density function for the number of successful attacks is given for days 10, 20, 30, 40, and 50. The probability densities for days after day 50 are not shown since they were very similar to that of day 50.
Note that on the 10^{th} day the probability that there has been a successful attack is less then 12% and that the probability that there have been more than one successful attack is less than 1%.
At day 20 the probability that there has been no event is already less than 37% and the probability that there has been one successful attack is also slightly less than 37%. Already the probability that there have been two or more successful attack has grown to a figure just under 27%.
By the 30^{th} day the probability that there has been no attack is only a mere 15%, whereas the probability that there have been one or two successful attacks is 28% and 27%, respectively. The probability that there have been six or more successful attack has now grown to a staggering 1%.
On day 40, the probability that there have been no attacks has decreased even further just under 10%. The probability that there has been one, two or three successful attacks is now an approximate 70% and the probability that there have been more than six events has grown to over 3%.
The probability distribution for day 50 is not so different from day 40. The probability that there has been no event has dropt another percent to just under 9%. But the probability that there have been multiple incidents has slightly risen.

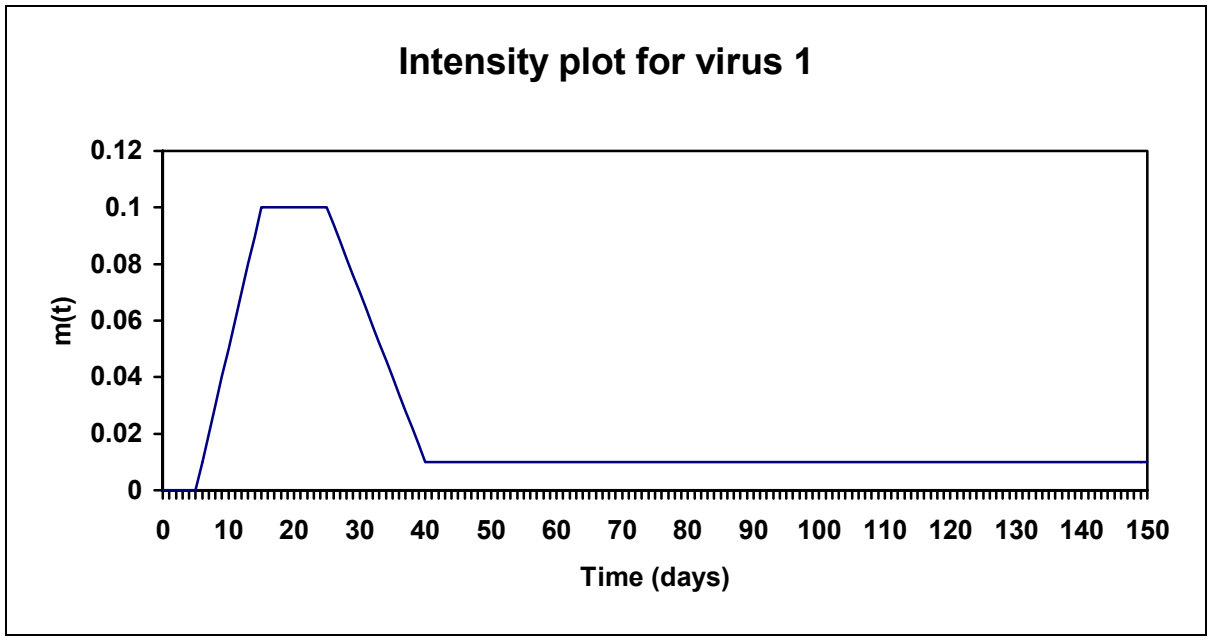

Figure 7: The linear intensity plot for virus 1 over 150 days.

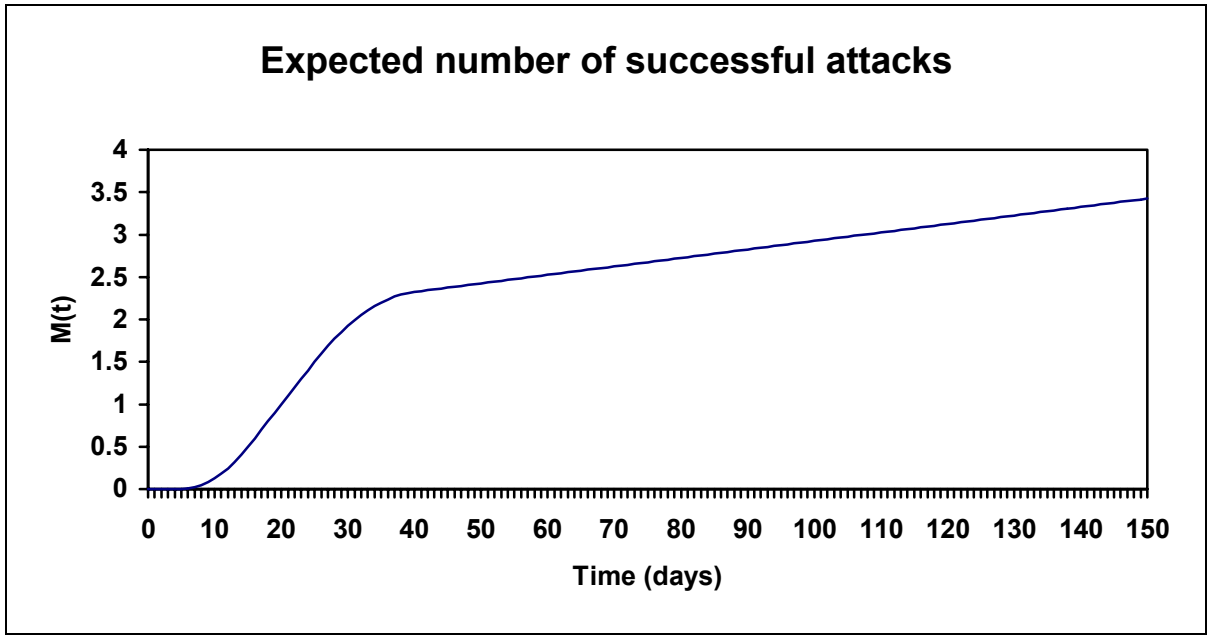

Figure 8: Expected number of attacks for virus 1 over 150 days.

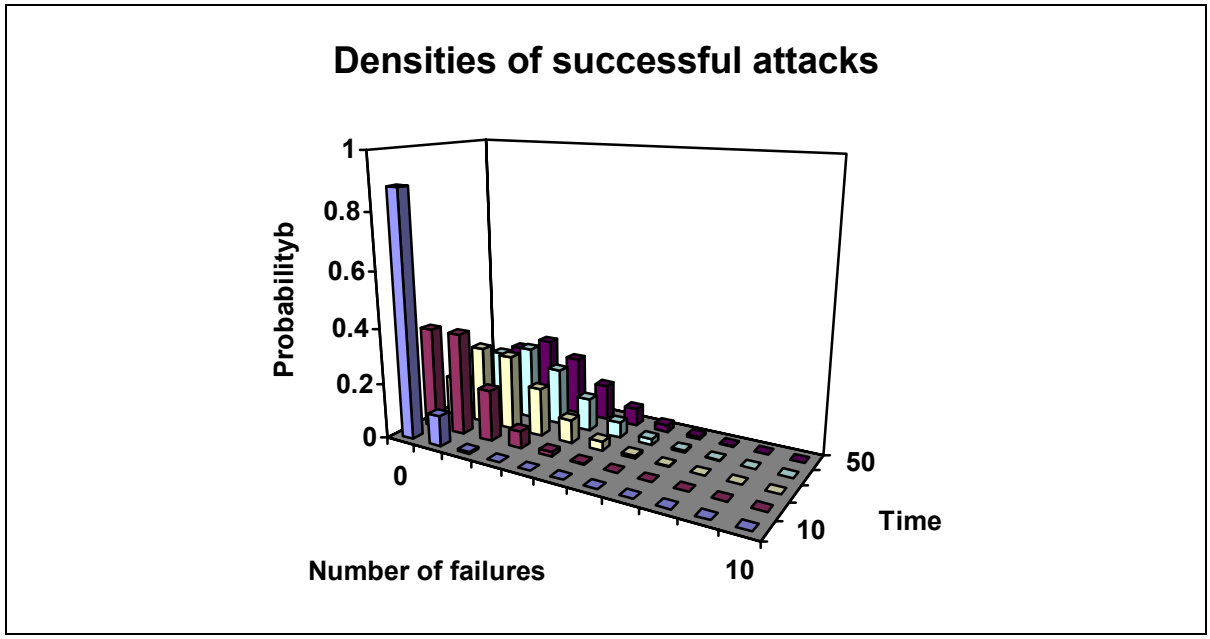

Figure 9: Probability distribution for the number of successful attacks over time.

3.1.4 The modeling of safeguards for viruses

Now that the general model for the hazard of attacks is discussed, in this section the safeguards will be incorporated into the model. From experience I have noticed that installing an antivirus, firewall and programs such as Spybot on your personal computer helps to keep it running properly. It lessens the chance of viruses or other malware settling on your computer. But because of the constant development of new viruses these programs always recommend updating every week or even every day.

But what exactly makes the difference between updating right away or somewhere next week is hard to say. Sure the companies advise to do it as often as possible but is that really useful and even necessary?

In this chapter, a model is created for these various safeguards as concomitant variables that influence the hazard, and, in addition the effect of different times of updating are compared. To this extent a log-linear model is proposed for the intensity of attacks. Denote the intensity of successful attacks of type i for protected systems type j with covariates x_{ij} by $\lambda(t,\beta_i)$, and define it as follows

$$\lambda_{ij}\left(t,\beta_i\right)=\tilde{\lambda}_i\left(t\right)e^{\beta_i^T x_{ij}(t)} \tag{3.2}$$

where the coefficients of the covariates β_i and the covariates x_{ij} are vectors of length m, explicitly

$$\begin{aligned}\beta_i^T &=\left(\beta_{i1},\ldots,\beta_{im}\right)\\ x_{ij}^T\left(t\right)&=\left(x_{ij1}\left(t\right),\ldots,x_{ijm}\left(t\right)\right).\end{aligned} \tag{3.3}$$

As said before, the covariates represent the various possible safeguards a system could have. Typically at the time of initiation of the i-th virus, t_{i1}, the safeguards have no way to protect systems against the new virus, but as the virus is detected and patches are created the viruses have less chance to do harm to protected systems. Assume that at time s_{jk} safeguard k comes up with a patch against virus i. In this case the covariates x_{jk} are indicator functions defined as

$$x_{ijk}\left(t\right)\begin{cases}0:t<s_{jk} \text{ or safeguard } k \text{ is not present on system } j\\ 1:t>s_{jk} \text{ and safeguard } k \text{ is present on system } j\end{cases} \tag{3.4}$$

The coefficients β_i can be interpreted as the effect safeguards have on the reduction of the hazard. Note that a value for β_i close to 0 yields a safeguard that is providing little protection and a safeguard with an extreme small value for β_i yields a safeguard that provides exceptionally good protection. Positive values for β_i on the other hand indicate safeguards that could more appropriately be called unsafe guards.

3.1.5 A single safeguard against a virus

This example considers the same virus as was considered in the first example. There is however the difference that a single safeguard is available. Furthermore, the time when the safeguard can be patched to make it resilient to the new virus is still open and different possibilities are explored.
In this example, an antivirus was chosen as it is a common safeguard against viruses. Because only one type of hazard is used, the subscripts *i* are dropped. The antivirus is modeled as a covariant with accompanying coefficient $\beta = -6$ such that it offers strong protection against virus attacks. Again all calculations were done using Excel.
Three different times for the proposed patches are compared. The first patch-time is halfway between the time that the virus has initiated and the time the use is saturated. The second patch-time is in the middle of the point of saturation and the point where the number of hosts for the virus start to decrease and the final proposed patch-time is halfway between the decrease of hosts point and the point where the intensity of the virus is for the first time at its minimum.
Summarizing, the 10^{th}, 20^{th} and 32.5 day as different patch-times for the antivirus are used. How these choices affected the hazard rates and the expected number of events is shown in Figure 10 and Figure 11, correspondingly. Notice that the intensities follow the original ones until the patch is added and that they then drop to a small value in comparison to the original value. The expected number of failures also follows the original expected number of failures until the antivirus is patched. From then on the expected number of failures does not seem to change significantly.
To compare how the safeguards are influenced over time the probability distributions of the number of successful attacks were calculated and the appropriate figures are supplemented, see Figure 12, Figure 13, and Figure 14.

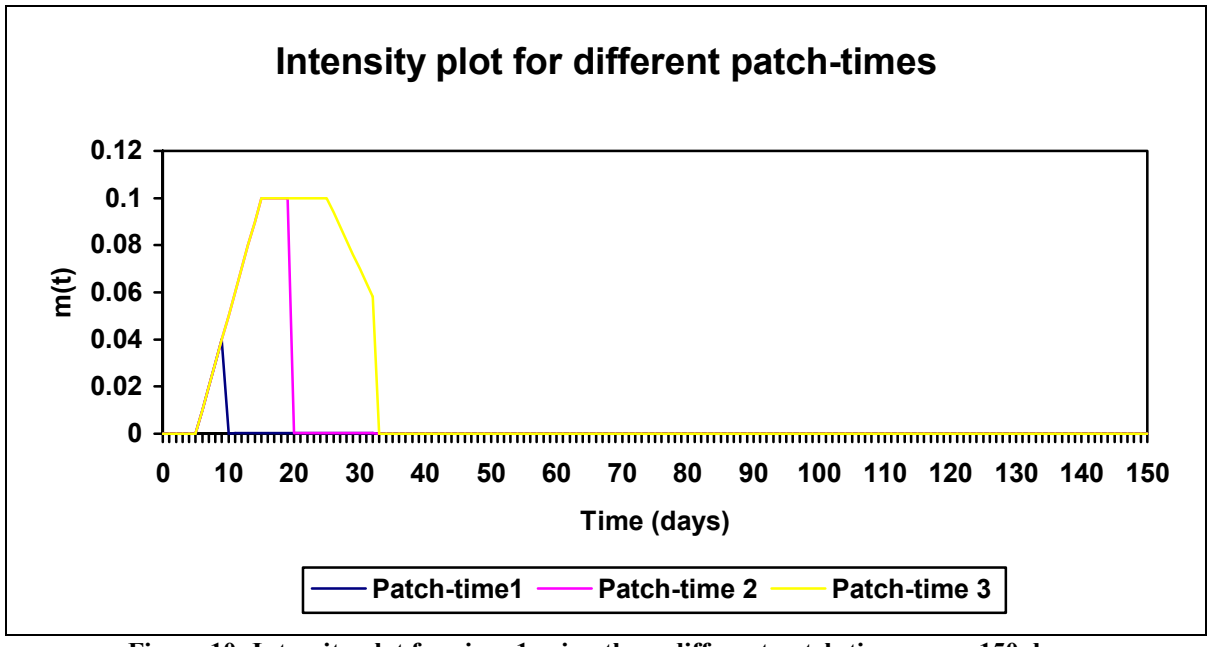

Figure 10: Intensity plot for virus 1 using three different patch-times, over 150 days.

From Figure 12 it is inferred that due to the patched antivirus after the 10th day, there is still a significant probability of a successful attack before then. In this case there was an approximate 11% chance that there has been one event. The probability that there has been more than one event was less than 1%. This left an approximate 88% probability that there has been no event in the first ten days.
Because the antivirus was patched at day 10, it is seen that the probability densities for the number of failures after that day do not change significantly. This can be explained by the strong protection offered by the antivirus.
In Figure 13 it is shown that at the 10th day the same probability distribution is obtained as for patch-time 1. The difference is that the probability distribution changes for the 20th day. Where there now is only an approximate 37% chance that there has been no attack and also an estimated 37% chance of one attack. Furthermore, the chance that there have been more than one successful attack is around the 26%. Note how the numbers correspond with the numbers obtained from the model with no safeguards. This is due to the fact that the safeguards do not have any backdated effect. From the 20th day onwards however the probability that there will be an additional attack is very slim due to the strong protection of the updated antivirus against that particular type of virus.
In Figure 14 the same result is noticed. The probability distribution for the number of viruses now follows the original probability distribution without the safeguards for the first 30 days. As a result the estimated probabilities stay the same.

Finally it is noted that other covariates can be incorporated in the same way. The amount of use of your system, for instance, may well influence the susceptibility to attacks. In this model another parameter could be created such that the intensity for moderate users is less because of the relative little exposure.

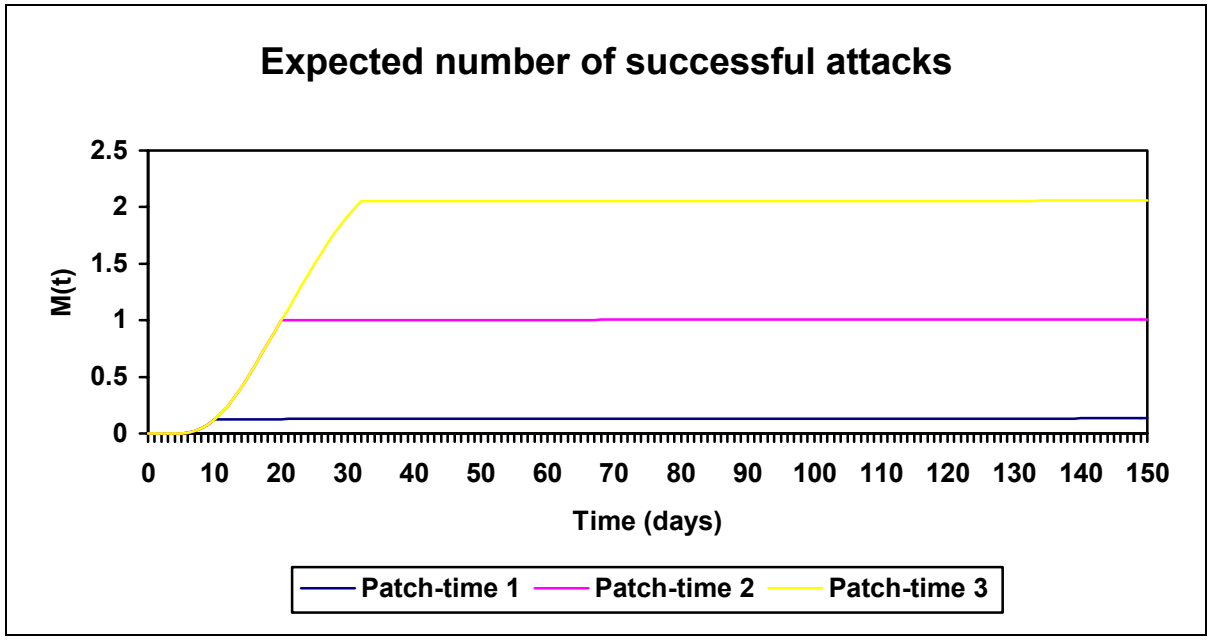

Figure 11: Expected number of attacks for virus one using three different patch-times, over 150 days.

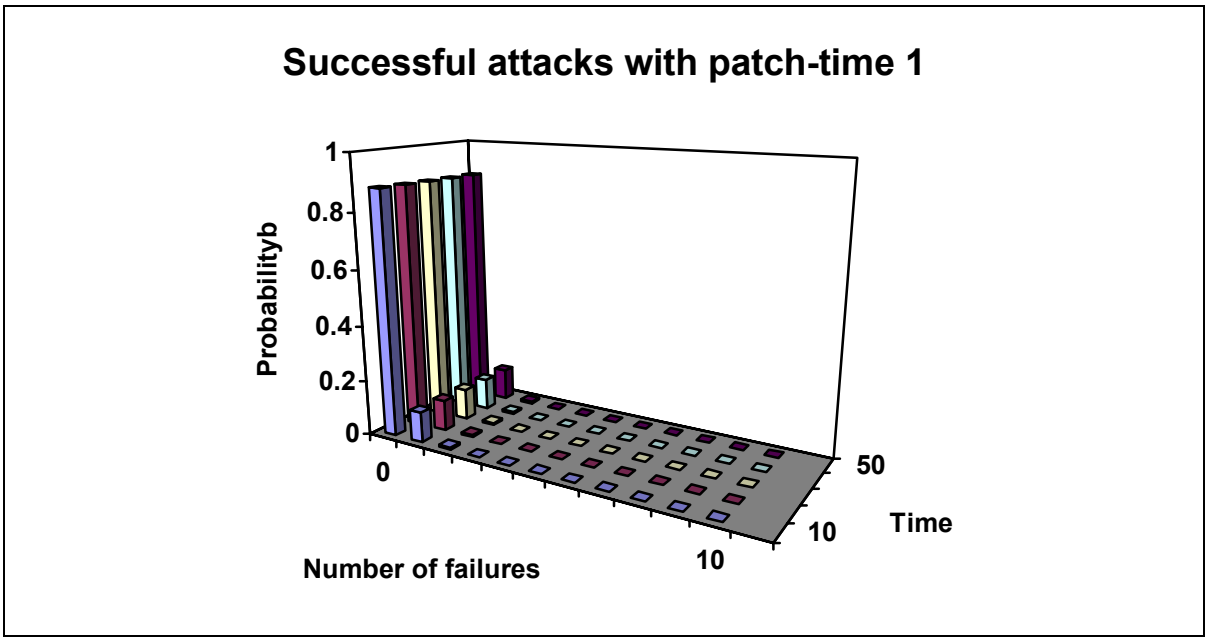

Figure 12: Probability distribution for the number of successful attacks for the first 50 days with a 10 day interval for patch-time 1.

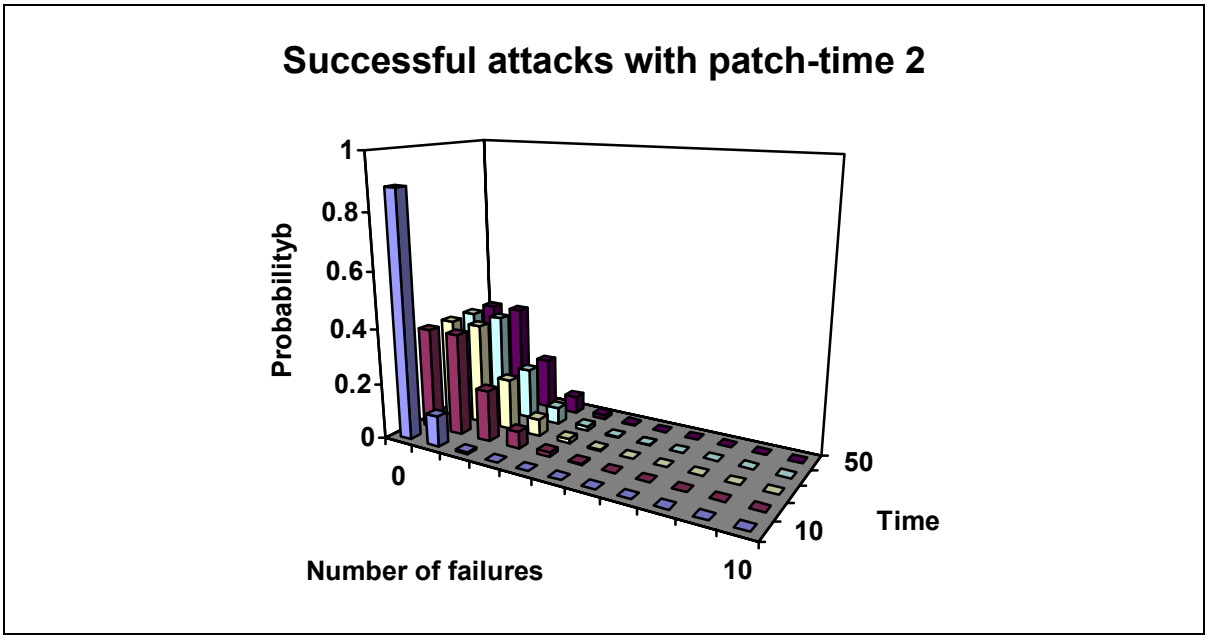

Figure 13: Probability distribution for the number of successful attacks for the first 50 days with a 10 day interval for patch-time 2.

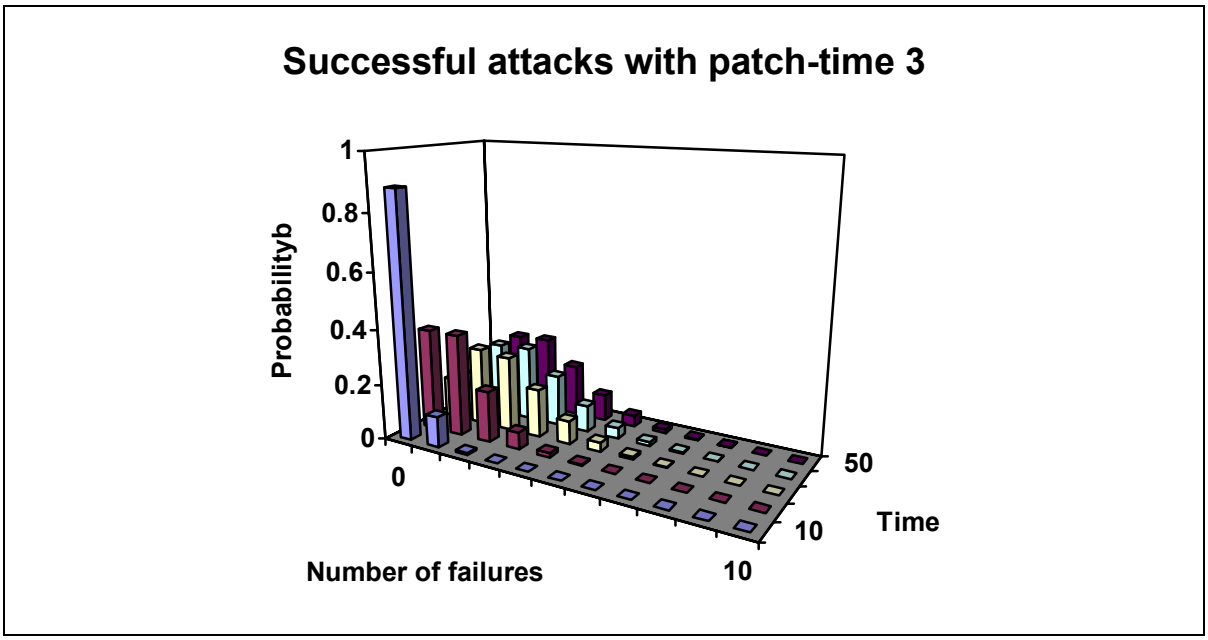

Figure 14: Probability distribution for the number of successful attacks over for the first 50 days with a 10 day interval for patch-time 3.

3.2 Securing Information Characteristics

In this section it is still assumed that the number of attacks at any time is Poisson distributed with a time varying intensity. However, in Section 3.2 it was chosen to model a safeguard that could counter the single type of attack. But in general this need not be the case, think for instance of the loss of more general information security notions such as the confidentiality, integrity and availability of information. Surely these information characteristics could not be affirmed with only one single form of protection.
As confidentiality is conceptually one of the most lucid information characteristics, it was chosen to consider it in more detail. To guarantee confidentiality however is generally not straightforward in the information security area. And, as a result, it was chosen to model the security of this information characteristic with respect to multiple safeguards.

3.2.1 Modeling confidentiality

Recollecting, confidentiality was the property that information access and disclosure is restricted to authorized personnel. The number of confidentiality breaches at every point in time *t* are considered Poisson distributed. Reasons for confidentiality breaches to occur are wide ranged. One could think for example of confidentiality breaches because of a disk being lost with confidential information, or a former employee being disgruntled over being fired and decides to spread confidential information or simply the disposal of a confidential document in the thrash that ends up in an office of a journalist.

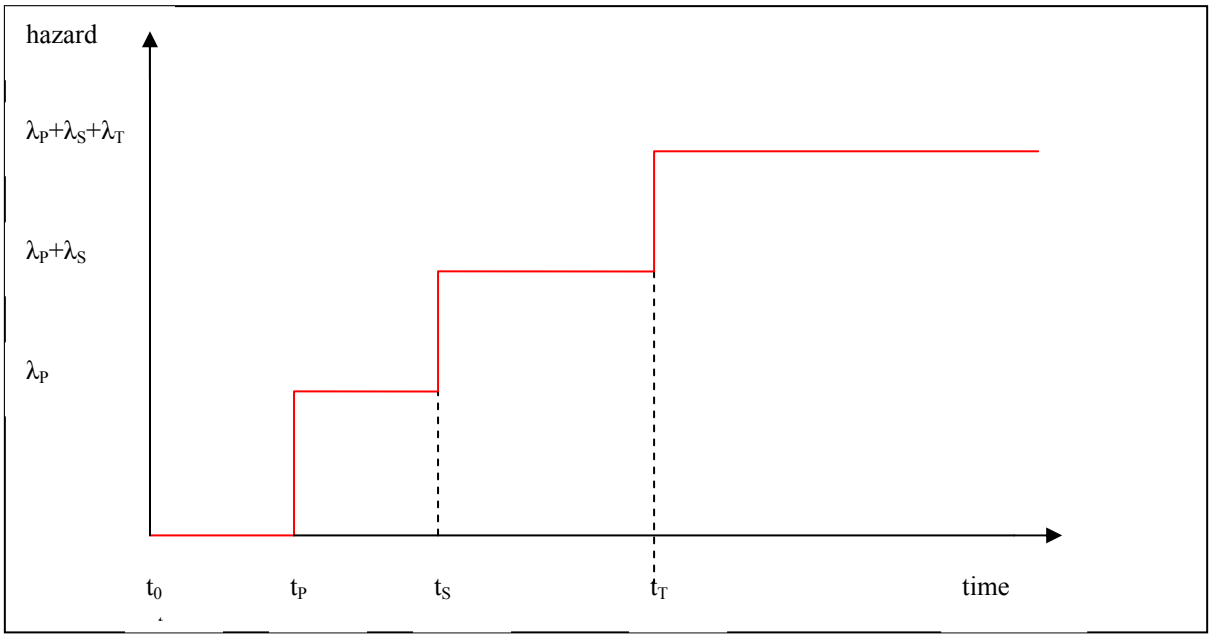

Figure 15: Illustration of intensity of the confidentiality attacks.

The fictitious confidentiality breaches modeled in this section are supposed to come about in three independent information states, namely storage, processing, and transmission that were previously discussed in chapter 1.
That the intensities in the three information stages are independent might be an oversimplification of the actual problem, but it enables one to investigate the complex notions and leads to interesting results.
The intensities of the number of successful attacks for information states storage, processing and transmission are assumed to be constant over time say λ_S, λ_P and λ_T and start at time t_S, t_P and t_T, respectively. An example of such a general intensity is given in Figure 15.
Note that in the case that the initiation times of the attacks on the different information states are the same the hazard becomes a step function with as intensity the sum of the three intensities. In this case the process is a superimposed Poisson process.

3.2.2 Modeling safeguards for confidentiality

The safeguards that can be employed for the security of confidentiality are the controls and countermeasures from chapter 1. Recall that the three categories in which controls and countermeasures could be divided were technology based, management based and mitigation based countermeasures.

The difficulty now is to model the effect the different safeguards have on reducing the intensities of the number of successful attacks for each different information state. In reality each different countermeasure from each different group may influence the intensities of the information states differently. Therefore it was chosen to use three different safeguards: a technology based, a management based and a mitigation based.

The technology-based countermeasure that was chosen is cryptography that as was said may well be one of the earliest recognized countermeasures for providing information security. In general, cryptography can be employed on two of the three information states, storage and transmission. It is however by definition impossible to use cryptography in the processing of

information since cryptography is used to make information unintelligent. Recall that in this context by the processing of information the use of information is intended. And surely one cannot use unintelligent information.

The management-based countermeasure that is selected is the clean desk policy. A policy in which each person is expected to clean their desks during absence. The clean desk policy is supposed to affect the physical storage of information, but is unlikely to affect the information states like transmission and processing.

Lastly, the mitigation based security measure that is investigated is training and awareness. This security measure is supposed to influence the processing information state. The processing for instance may be secured because of the awareness of employees that some information is confidential and must be kept within a company.

3.2.3 Example: which safeguard to deploy first for maintaining confidentiality of Company A?

As an example, the safeguards mentioned in the previous section cryptography, the clean desk policy and training and awareness, and in what way they may secure the information characteristic confidentiality of Company A have been modeled in this subsection.
The model assumes that the first day in fact was the day the company was started, and so, this is proposed to be the initiation time for the hazards. And, hence, it is the day when the problem of securing confidentiality for Company A started. Already from the start Company A has been aware that information security needs to be addressed at all times. And as a result already from the opening the company starts implementing safeguards to protect their information.

One question that Company A still needs to address is the following: is there a specific order in which the safeguards could be implemented such that the expected number of successful attacks is minimized?
Questions like this are generally difficult to answer, but in the example below it is shown that with some extra information, this question could be answered under model assumptions. And, moreover, the succession in which the safeguards should be implemented such that hazard is minimized can be specified.

Now assume that the attacks on Company A's information confidentiality are threefold, reflecting the storage, processing and transmitting phases of information. Assume furthermore that the intensities are given by $\lambda_S = 0.25$, $\lambda_P = 0.14$ and $\lambda_T = 0.06$.
Besides these intensities, there are supposed to be three different options for implementing safeguards, namely cryptography, the clean desk policy and the training and awareness. Because information safety was one of the top priorities for Company A all these safeguards needed to be developed or bought and implemented. But due to the fact that the development/buying and implementation of each safeguard was time consuming they had to be adopted successively.

The research department gave the following estimations. For the development and implementation for cryptography, the time till development and implementation will be 40 days. But it then offers strong protection against transmission with accompanying coefficients of the covariates $\beta_{TI} = -5$.

For the development and implementation of the clean desk policy, the estimated time for implementation is 20 days. It offered good protection against storage, that is to say the accompanying coefficient of the covariate was $\beta_{S1} = -3$.
And the assessment for time needed for the training and awareness to come into affect was estimated by 90 days. The training and awareness is supposed to offer strong protection in the processing state with accompanying coefficients $\beta_{P1} = -4$.

With this additional data, the answer to the decision question in which order, if any, the safeguards needed to be implemented such that the number of successful attacks was minimized can now be answered.
Since the probabilities of a successful attack are supposed to be Poisson distributed for every choice of order for the safeguards, and the Poisson distribution is a function of the cumulative hazard, the order for the safeguards that minimizes the cumulative hazard is the order that minimizes the expected number of successful attacks.

Table 2: The cumulative hazards for the ordering of the implementations of the safeguards

Order of safeguards	Cumulative hazard
TSP(123)	27.77
TPS(132)	38.57
STP(213)	27.42
SPT(231)	47.41
PTS(312)	58.56
PST(321)	58.2

In this case with only three different safeguards that results in 6 different possibilities for the order of implementation for the safeguards. Calculation of the order that minimizes the expected number of attacks could be done by hand, but it is noted that more complex problems could be solved by linear optimization.
All different orders in which the safeguards could be implemented were checked and the results can be found in Table 2. As can be seen the order of the safeguards that minimizes the cumulative hazard is to implement the clean desk policy first, the cryptography second and the training and awareness last. This is somewhat surprising since the first safeguard that is implemented does not address the highest component of the intensity from the information state. It may well be due to the fact that the implementation of this safeguard only takes 20 days, so that by implementing it first already one of the components of the intensity is secured.
Furthermore, the difference between choosing to safeguard first the transmission phase, second the storage phase and lastly the processing stage and the optimum solution is small and only leads to an increase of an approximate 0.36 in the hazard. Although this does not seem like much it may well be the difference between an actual successful attack and no attack at all.

The figures for the individual intensities and the expected number of successful attacks for the optimum implementation order of the safeguards are given in Figure 16 and Figure 17.
Note that minimizing the cumulative hazard is the same as minimizing the area under the lines of Figure 16. It is seen that the order depends on the estimated amount of time it takes for the safeguards to come in effect, the size of the estimated intensities, and the level of protection.
In Figure 17 the contributions of the different type of attacks to the total expected number of successful attacks are given as a function of time.

In Figure 18 the shift for the first 50 days of the probability distribution for the number of successful attacks are shown. Up to day 20 the probability distribution for the number of successful attacks grows swiftly and seems to decelerate a bit after the 20th day. This is due to the storage safeguard comes into affect on that day.
From Figure 19 it can be inferred that the probability distribution for the number of failures does not seem to grow that fast after day 60. This is due that on the 60th day the safeguard for storage is implemented and that from that day the two information states with the relative biggest hazards are protected. This trend develops over the next 90 days up to the 150th day when the processing safeguard comes into affect. At this point in time the hazard is increasing marginally since all the information phases are protected.

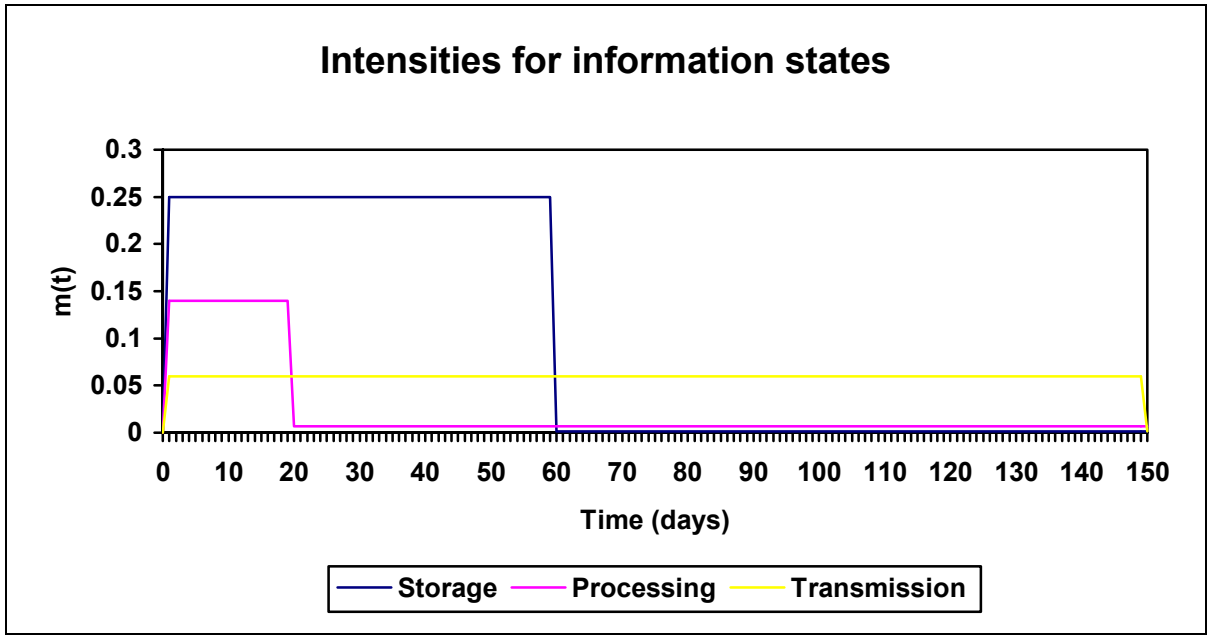

Figure 16: Intensities for the information states storage, processing and transmission.

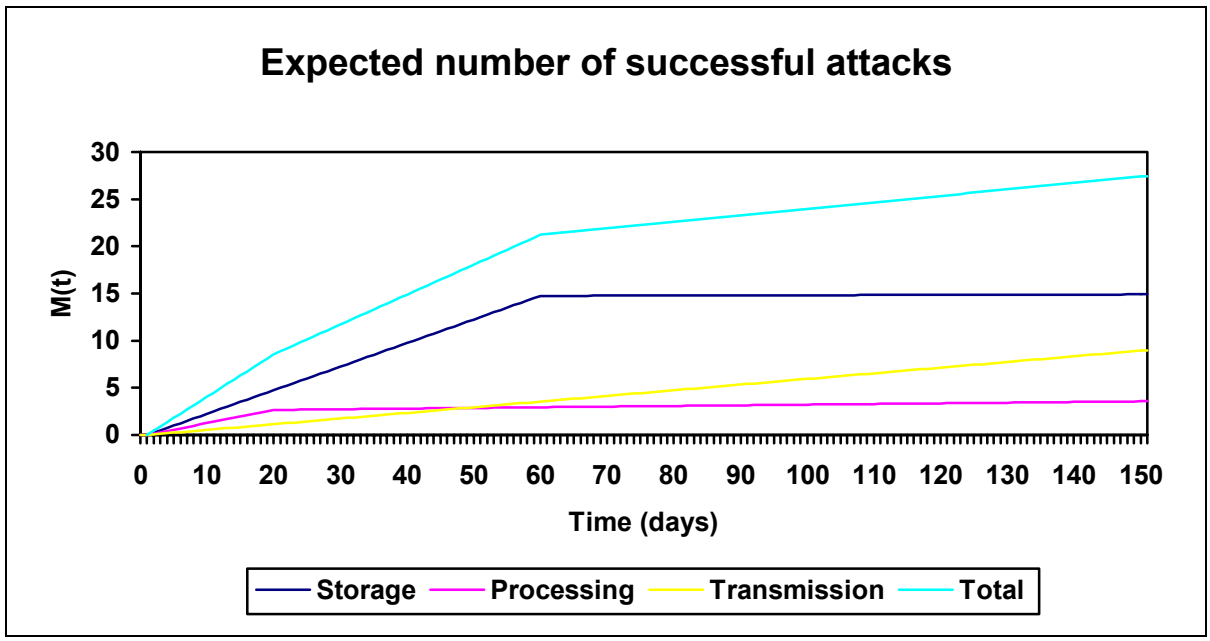

Figure 17: Expected number of successful attack from the different information states and the total number of successful attacks.

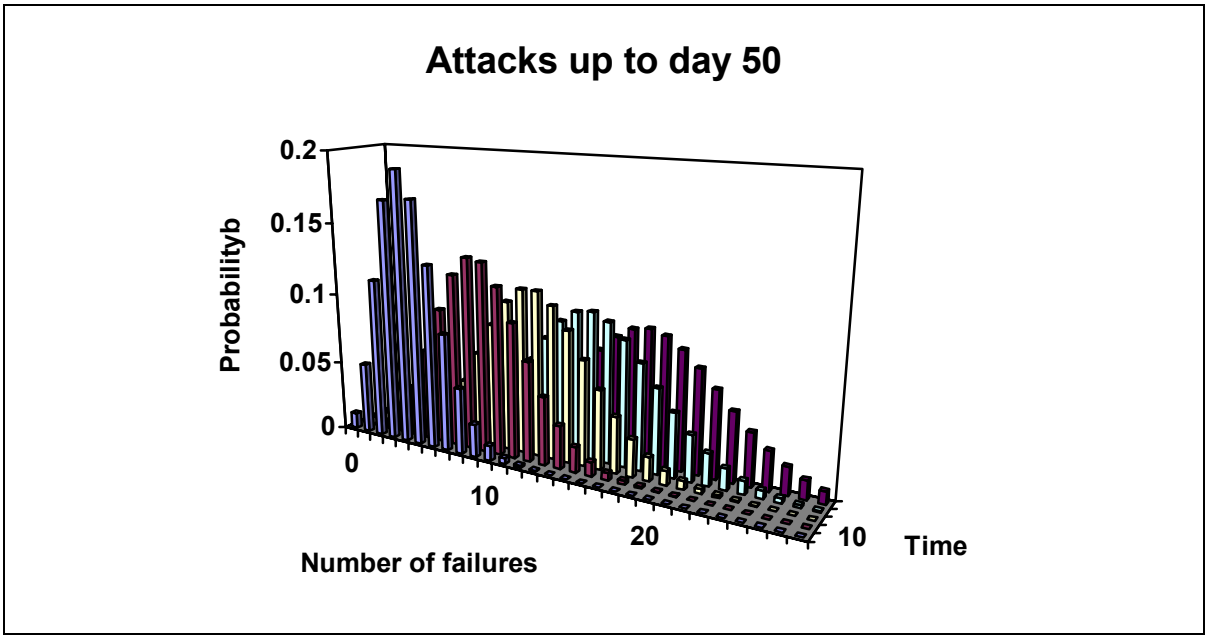

Figure 18: Probability distribution for the number of successful attacks for time up to day 50 with a 10 day interval when the optimum order of safeguard implementation is chosen.

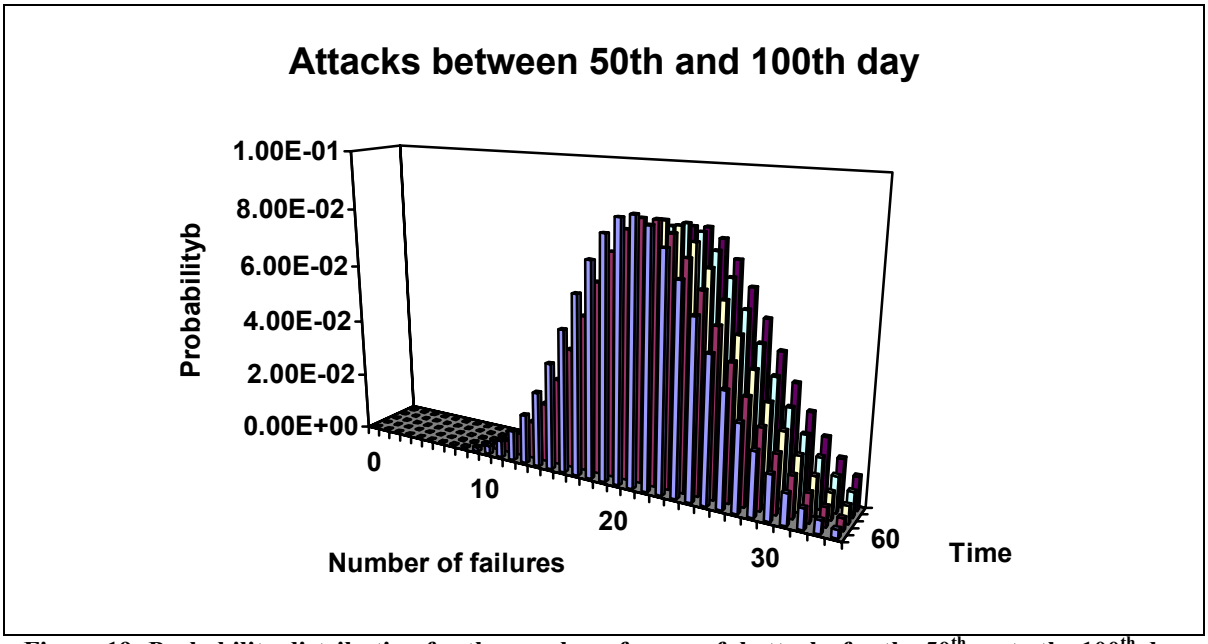

Figure 19: Probability distribution for the number of successful attacks for the 50th up to the 100th day with a 10 day interval when the optimum order of safeguard implementation is chosen.

3.3 Conclusions and recommendations from chapter 3

In this chapter it was found that the modeling of the number of successful attacks with the Poisson model has lead to insight into the use of safeguards. In both practical and more theoretical cases it was found that this model gives solid quantitative grounds upon which information security decisions could be based.

By considering the intensity of a successful virus attack as piecewise linear it was found that swift patch-times could have a significant contribution to reducing the expected number of attacks. That is to say, the model indicates that user vigilance at all times can seriously downsize the chance of malicious infections. Since this model only pinpointed one particular attack, it may well be possible to model other malicious infections in similar ways as to obtain further insight in the information security mechanism. Additionally, the model could be adapted to a particular form of virus propagation to test for model accuracy. Also, the effect of the incorporation of other safeguards could be interesting.

Furthermore, it was found that the expected number of successful attacks can be influenced by investigating different safeguard properties such as implementation length and comparing relative intensities. With the Poisson model assumption it was established that different implementation orders influence the expected number of successful attacks significantly.
A drawback of this model might well be that expected losses were not calculated. The reason to refrain from incorporating expected losses is that estimation of the expected losses is difficult, if not impossible. And so the choice not to incorporate the expected losses was done on the assumption that the expected losses may well be of the same magnitude over the entire width of the information security spectrum. However, if expected losses could be calculated accurately it is the writer's opinion that it is not impossible to incorporate those in the model to make decisions based on the minimization of expected costs.
In addition, it would be interesting to calibrate model parameters to real data. To this extend for instance maximum likelihood techniques could be used. When calibrating model parameters the interesting topic of inference would have to be addressed, which I have been able to avoid so far. This was done since the role of inference in other models is already addressed extensively in Chapter 4 and Chapter 5 of this thesis.

The chapter is finished by concluding that the information security sector is an interesting new area for quantitative research. Because we only just entered the information age and there is no indication that this is ending any time soon, it is also my belief that new topics will crop up in due time.

4 Applying Proportional Hazard to Information Security

This chapter deals with one of the main problems in the field of reliability: the dealing with censored data. First of all, the gathering of reliability data is often extremely expensive. And second, the data though censored still gives valuable information about system performance. Because of these facts it would be imprudent to simply ignore the censored data.
Cox's Proportional Hazard Model (PHM) is a model that is able to incorporate censored data and is therefore widely used in the field of reliability, ranging from biostatistics to engineering.
In this chapter the PHM will be further investigated as it is a way to deal with data obtained from many different Poisson processes, that is to say every datum is considered to be the result of a particular Poisson process, which stops after the first failure.
The purpose for the use of data in this chapter was to sketch the practical difficulties of working with reliability data.

This chapter considers networked information systems. The systems are divided into two different groups: the first group consists of systems that do not have any protection worth mentioning. The second group of systems on the other hand has an enhanced state of security, that is to say they have up to date firewalls, anti-viruses, etc.
From a system administrator's perspective the second group should have fewer information security breaches, since hackers and viruses have, luckily for the system administrator, relative more trouble getting into a heavily secured system than an unsecured system. If there only was one type of these protection systems one could imagine that the system administrator would not have much trouble making his decision, since a protected system outperforms an unprotected system. The catch however is that software is not free, and , furthermore, from every type of software one has the choice out of several brands. This makes the system administrator's job somewhat more difficult since they generally will have to decide if the extra protection is worth the cost, and, if so, what the best level of protection given a certain maximum investment is. In order to make justifiable choices the system administrator would like to quantify the benefits of the different systems. Ideally one likes to relate an attack to an amount of loss expressed for example in currency, this yet is not straightforward since one first has to estimate the expected loss of an attack. Ryan and Ryan [2005:1] proposes the use of expected benefit as a way to determine the return on investment (ROI). This involved among others the estimation of probability of successful attacks and the outcome of the attack.

My choice in this chapter has been to use the process of baselining. According to Whitman and Mattord [2003] a baseline is a "value or profile of a performance metric against which changes in the performance metric can be usefully compared". It was chosen to model the distribution of the time till failure from the different systems, which is theoretically more simplistic than the ROI strategy but has the advantage that the results can be easily interpreted since the distribution of the time till failure can be used as a baseline and changes can be assessed on that basis. To this extend Cox's widely celebrated proportional hazard model (PHM) is used.

First the dataset that will be considered is given in section 4.1. In section 4.2 the survival analysis framework is given and survival curves are computed. In section 4.3 traditional hypothesis tests are performed and in section 4.4 the PHM is explained. In section 4.5 a cause specific model is investigated. And, finally in section 4.6 conclusions and recommendations are given.

4.1 The data

Ryan and Ryan [2005:2] apply the PHM to a group of systems that are studied for a period of 100 days. The data is initially subdivided into two groups: the first group for which the systems have no noticeable protection and the second group which has good protection, see Table 3 and Table 4. Furthermore the type of failures have been registered and subdivided into the appropriate critical information characteristic, i.e. availability, confidentiality and integrity as they are used by for instance McCumber [2005].
It is noted that this is simulated data. As a result even from censored observations the cause was known. In reality, however, the cause is by definition unknown. This however does not undermine the goals of this chapter, as it was intended to show the practical difficulties of handling censored data.
The data points are times and were supposed to be checked every 12 hours. Minus signs indicate the time at which censoring of the system was noted. *C*, *I* or *A* indicate the cause of the failures, confidentiality, integrity and availability, respectively.

The first group is the group of unprotected systems, which are frequently attacked and after 50 days, see Table 3. As an example the first cell in the table indicates a integrity failure at time *0,5*, and the cell indicates a censored integrity failure at time 100. It is noted that in general the causes of a censored observation are unknown, but this slackness does not influence the results of this chapter.
From the table it is seen that already at half the study period 80 percent of the systems has been successfully attacked and at the end of the study only 6 percent of the systems functions adequately.

Table 3: Notional computer system failure for unprotected systems, data in days and - sign indicates censoring. The causes of failure are A = availability, C = confidentiality and I = integrity.

0,5 I	0,5 I	19,5 A	20 I	23,5 I	23,5 I	25,5 C	25,5 C	-27,5 A	27,5 I
27,5 A	28,5 I	-29 I	29 A	29,5 I	30 I	30,5 A	30,5 I	31,5 I	33 I
33 A	33 C	33,5 C	33,5 I	34 I	34,5 I	34,5 I	34,5 C	35,5 I	35,5 A
-36 I	36 I	37 I	37 I	37 A	37,5 A	37,5 C	38,5 I	38,5 I	38,5 I
38,5 C	39,5 I	39,5 I	39,5 C	40 I	40 C	40,5 I	40,5 C	40,5 I	-41 A
41 I	42 I	42 A	42,5 A	42,5 I	43 I	43 I	43 I	43 C	44,5 A
44,5 A	44,5 A	44,5 I	-45 I	45 A	46 A	46 I	46 I	46,5 I	46,5 A
47 I	48 I	48,5 I	48,5 I	48,5 C	48,5 I	49 I	50 A	50 I	50 I
-51 C	51 I	52 I	52 I	-53 I	53 I	53,5 I	54,5 I	55 I	56 I
57 I	57 I	59,5 I	61 I	-100 A	-100 A	-100 I	-100 I	-100 I	-100 I

The second group consists of protected systems, which can be thought to be more reliable, see Table 4. After 50 days of study only 3 percent of the systems have been attacked successfully, after this time the protected systems also show vulnerability to attacks. From this study group nonetheless over 30 percent of the systems remain to function adequately till the end of the study.

Table 4: Notional computer system failure for protected systems, data in days and - sign indicates censoring. The causes of failure are A = availability, C = confidentiality and I = integrity.

0,5 A	5,5 I	26,5 I	57 I	57,5 I	60 I	60 I	62 I	62 I	-65,5 I
65,5 A	67,5 I	68,5 A	70,5 I	72,5 I	72,5 A	72,5 I	72,5 I	72,5 C	75 I
75,5 I	77 I	77 I	78 I	78 C	79 A	80 I	80 I	80,5 I	81 A
81 I	82,5 I	82,5 I	84,5 A	84,5 I	84,5 C	85,5 I	85,5 A	86 C	86 C
87,5 A	87,5 A	87,5 I	89 C	89 I	89 C	91 I	91,5 I	91,5 I	91,5 C
93 I	93 I	-93,5 C	93,5 A	94,5 I	94,5 I	-96 A	96 I	96 I	97,5 C
98 I	98 I	99 A	99 I	99 I	99 I	-100 I	-100 I	-100 A	-100 C
-100 I	-100 I	-100 A	-100 C	-100 A	-100 A	-100 I	-100 I	-100 I	-100 A
-100 C	-100 I	-100 C	-100 I	-100 A	-100 I	-100 I	-100 I	-100 C	-100 I
-100 I	-100 I	-100 I	-100 A	-100 C	-100 I	-100 C	-100 I	-100 A	-100 I

So, after having observed the tables our intuition already leans toward the fact that the protected group is less liable to attacks than the unprotected group. Question remains however if the difference is statistically significant and if so how much better it is.

4.2 Survival analysis framework

Let $x_{1,1}, \ldots, x_{n1,1}$ and $x_{1,2}, \ldots, x_{n2,2}$ denote the ordered sample values belonging to the unprotected and protected systems, respectively. Then, if there are no censored observations, a convenient way to represent and visualize the data is the estimated cumulative density function (CDF) of the failure times, defined as

$$\bar{F}_n(t) = \frac{\#x_i \le t}{n}. \tag{4.1}$$

The CDFs could then be used to visually test if the two sample distributions vary or are in fact not so different.
With the given data Equation (4.1) cannot be used because it is not able to deal with censored data. The estimator most commonly used is the product limit estimator or the Kaplan-Meier estimator introduced by Kaplan and Meier (1958). It is considered as a nonparametric estimator, but more appropriately it would be called an estimator with infinitively many parameters, since the estimator maximizes the likelihood for the survivor function when the survivor function is not restricted to a certain parametric form. The product limit estimator is given by

$$\hat{S}(t) = \prod_{j|t_j<t} \left(1 - \frac{d_j}{n_j}\right) \tag{4.2}$$

where d_j is the number of failures at time t_j and n_j is the number of objects still in the study prior to time t_j. In case of no ties and no censoring this estimator equals the CDF from Equation (4.1). It is noted that the product limit estimator is the "nonparametric maximum likelihood estimate", see Kalbfleish and Prentice (1980) for the derivation of the maximum likelihood estimator.

At every fixed point in time the estimator only gives a point estimate for the CDF. Already in the early 20th century Greenwood came up with an asymptotic variance for life table estimator. The estimator for the variance is known as Greenwood's formula, Greenwood (1926), and is given by

$$\hat{Var}\left(\hat{S}(t)\right) = \hat{S}^2(t) \sum_{j|t_j<t} \frac{d_j}{n_j\left(n_j - d_j\right)}. \tag{4.3}$$

It can be shown that Equation (4.2) is asymptotically Gaussian distributed so that a confidence interval for the survivor function can be computed. For fixed t the 95% confidence interval is given by

$$\left[\hat{S}(t) - 1.96\sqrt{\hat{Var}\left(\hat{S}(t)\right)} \quad \hat{S}(t) + 1.96\sqrt{\hat{Var}\left(\hat{S}(t)\right)}\right].$$

At extreme values this estimator that theoretically is restricted to the [0,1] interval, since it is a cumulative density, yields values that are not restricted to this range. This can be avoided by estimating the variance of a transformation of the CDF. The asymptotic variance of

$$\hat{v}(t) = \log\left(-\log \hat{S}(t)\right)$$

can be used yields an asymptotic variance estimated by

$$s^2(t) = \frac{\sum_{j|t_j<t} \frac{d_j}{n_j\left(n_j - d_j\right)}}{\left[\sum_{j|t_j<t} \log\left[\frac{\left(n_j - d_j\right)}{n_j}\right]\right]^2}. \tag{4.4}$$

Confidence intervals can be computed in a similar way as for Greenwood's formula. These confidence intervals are known to be confined to the [0,1] interval. In Figure 20 the product limit estimators and 95% confidence bounds are computed for the protected and unprotected systems with the aid of Equations (4.2) and (4.4). The calculations were done in Excel.

Figure 20 seems to visually confirm the suspicion that the survival curves from the protected and unprotected systems differ significantly. In addition, observe that the estimator for the variance from Equation (4.4) is not monotonic decreasing as is to be expected. See for example the lower bound of the protected system which even seems to increase for the first 58 days. This is technically not possible since the survivor function should decrease on the entire range. One explanation is that the extra failures over time make the variance estimator less

uncertain. Furthermore, there is a dramatic decrease of the lower bound for the protected system at day 72. This may be due to the "massive" number of failures at that specific day – percentage wise that is.

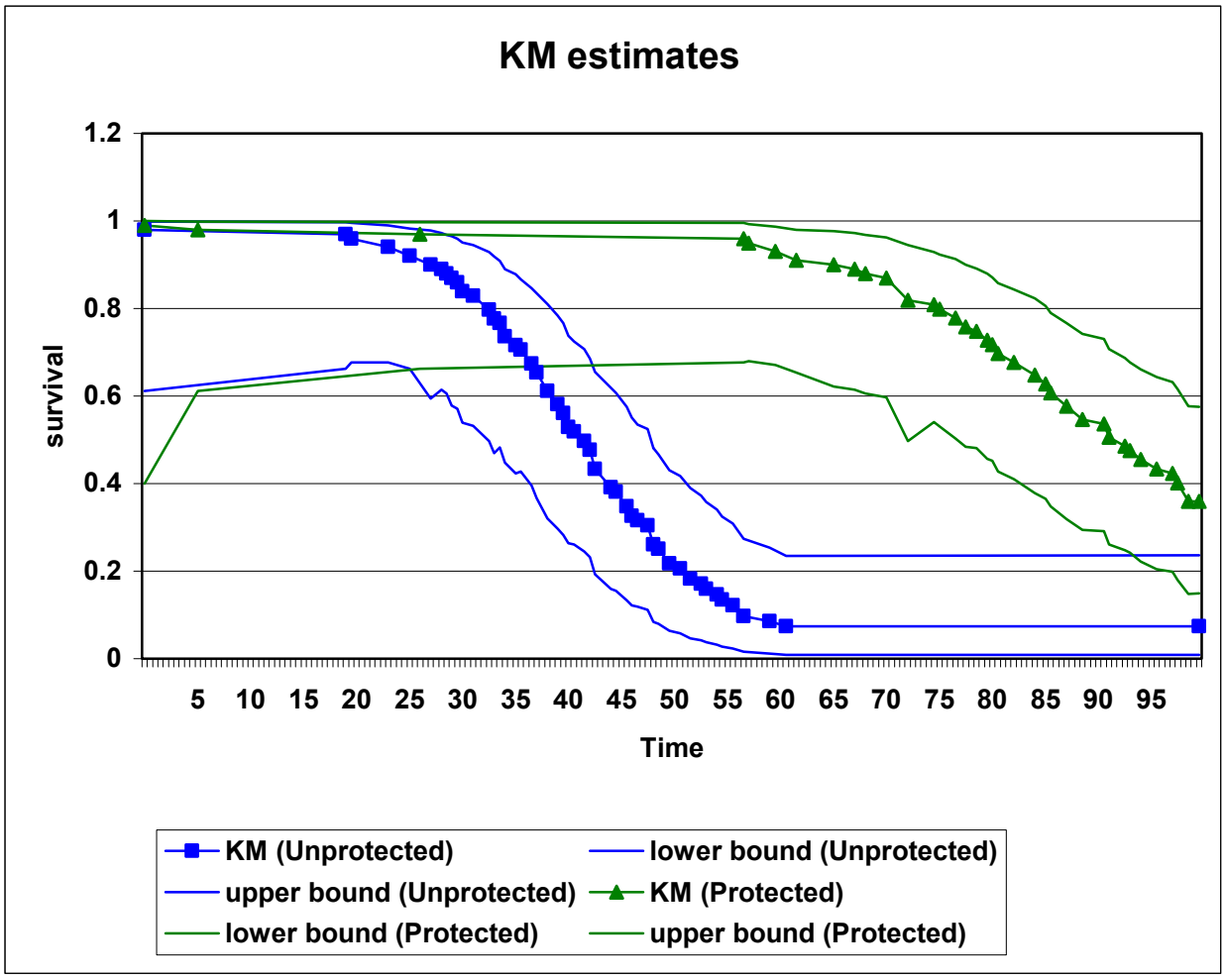

Figure 20: KM estimates for the protected and unprotected system with 95% confidence bounds.

4.3 Hypothesis testing for the survivor function

But still, though it is visually clear that the two estimated survivor functions from Figure 1 are not likely to have arisen from the same distribution function. One is not able to reject the hypothesis on basis of graphs. To this extent multiple test exist which enable to test for different survivor curves, for example the likelihood-ratio test, score test and log-rank test. In this example the log-rank test was chosen.
First of all, let's denote the ordered failure times by $t_1,\ldots,t_r$. Let d_j denote the number of failures that occur at time t_j and d_{ij} the number of failures at time t_j of type i. Furthermore, the total number of subjects at risk prior to t_j is given by n_j and as above n_{ij} the number of subject at risk prior to t_j of type i.
The joint-distribution for $d_{1j},\ldots,d_{rj}$ conditioned on d_j is given by

$$f_{d_{1j},\ldots,d_{rj}|d_j} \frac{\prod_{i=1}^{r}\binom{n_{ij}}{d_{ij}}}{\binom{n_j}{d_j}}. \tag{4.5}$$

The mean and variance for d_{ij} are known since Equation (4.5) is simply a hyper geometric distribution. The mean is

$$w_{ij} = \frac{n_{ij} d_j}{n_j}$$

and the variance is given by

$$\left(V_j\right)_{ii} = \frac{n_{ij}\left(n_j - n_{ij}\right) d_j \left(n_j - d_j\right)}{n_j^2\left(n_j - 1\right)}. \tag{4.6}$$

The covariance of d_{ij} with d_{lj} is

$$\left(V_j\right)_{il} = \frac{-n_{ij} n_{lj} d_j \left(n_j - d_j\right)}{n_j^2\left(n_j - 1\right)}. \tag{4.7}$$

Let v_j be defined as

$$v_j = \left(d_{1j} - w_{1j},\ldots, d_{rj} - w_{rj}\right)^T. \tag{4.8}$$

Then the statistic v_j has conditional mean zero and variance matrix V_j. Now define v as the sum of all v_j and V as the sum over all V_j then the test for equality of the survivor curves can be given by

$$v'Vv \tag{4.9}$$

This statistic follows a χ^2_{r-1} distribution with $r - 1$ degrees of freedom, since the elements of v sum to one.
The data from Table 3 and Table 4 gave the following values for v and V

$$v = \begin{pmatrix} 50.89 \\ -50.89 \end{pmatrix}$$

and

$$V = \begin{pmatrix} 23.39 & -23.39 \\ -23.39 & 23.39 \end{pmatrix}.$$

This yields a value of 110.70 as an approximation for the χ_1^2 statistic. Excel reported the significance as 6.88 E-26. Certainly with such a low value the null hypothesis that the data from Table 3 and Table 4 could have arisen from the same CDF by chance alone has to be rejected – by conforming to a significance level of 0.05. This confirms our earlier suspicions from Figure 20 that the survivor functions are significantly different.

4.4 Proportional Hazard Model

The proportional hazard model is well suited to model the effects of covariates on time till failure estimations. Cox (1972) proposed a new view on the parameter estimation, the partial-likelihood estimator that in the case of time-independent covariates can be interpreted as the likelihood of the order of the failures. Note that this is not an ordinary likelihood since it is not a product of probabilities of making the particular observations. But it is interpreted as the likelihood that the failures occur in that particular order. Breslow (1974) elaborated on this and showed that the partial likelihood in case of ties could be estimated by

$$L(B) = \prod_{d(t_i)>1} \frac{\exp\left(B^T s_i\right)}{\left[\sum_{t_j > t_i} \exp\left(B^T Z_j\right)\right]^{d(t_i)}}. \tag{4.10}$$

Where B are the coefficients of the covariates which are to be estimated, $d(t_i)$ are the number of failures at time t_i, Z_i is the vector of covariates belonging to even i and s_i is the sum of the covariates of individuals observed to fail at t_i. Note that in this example the covariate vector is taken one-dimensional and that it is of a dichotomous nature, i.e. zero if the system is not protected and one if it is.
Calculation of the maximum value of the coefficient was done with the Newton-Raphson algorithm. This is an iterative method for finding for example extremes of functions. It is based on the prediction of the gradient vector and the Hessian matrix. The algorithm consists of 2 steps, the prediction step in which the gradient and the Hessian matrix are updated and the maximization step in which the estimate for the coefficients is updated. Equation (4.11) gives the formula used to update the covariates. This is the standard Newton-Raphson algorithm for finding extremes and is known to converge for concave functions. Rathbun (1996) gives a clear overview of the algorithm and formulas for the score vector and Fischer's information matrix can be found in Kalbfleish and Prentice (1980).

$$\hat{B}_{i+1} \approx \hat{B}_i + \left(I\left(\hat{B}_i\right)\right)^{-1} \frac{\partial}{\partial B} \log L\left(\hat{B}_i\right) \tag{4.11}$$

The algorithm was implemented in Excel. Already after 4 iterations and initial estimate 0 for the coefficients it was found that the point value for maximizing the partial-likelihood in this example is given by $\hat{B} = -1.82$. This is intuitively clear since the sign indicates that an increase in the covariate decreases the hazard.
In the absence of ties the asymptotic distribution of $\hat{B}$ is proven to be normal with mean β and covariance matrix $I\left(\hat{B}\right)^{-1}$. Under the assumption that this result holds in case of ties the 95% confidence interval is given by *(-60.71, 57.34)*. Since the interval contains *0* we can not reject that the covariate does not have a positive influence.

Note that the main assumption of the model is that the hazards are proportional. This implies that the survivor functions can be written as roots of one another. Since this is difficult to check visually a figure of the estimated cumulative hazard is provided, see Figure 21. This figure also should show proportionality under model assumptions, because it is a sum of the proportional instantaneous hazards. This was actually proposed by Lawless in his part of the discussion on the paper of Ansell and Phillips [1989].
From the previous figure it is inferred that the hazards do not look proportional. Until day 20 the hazards seem to coincide, but past day 20 the unprotected hazard seems to grow exponentially, whereas the protected hazard stays low for the next 40 days. Moreover, after day 60 this hazard is also increasing rapidly.

4.5 Cause-specific hazard model

The model used in this section is based on the same observations as the one-covariate example. The difference is that cause-specific hazard functions are now considered as was described by Kalbfleish and Prentice [1980] and used on this problem by Ryan and Ryan [2005:2]. The difference with the one-covariate model is that now additional information about the failure times is used. In addition to the time of failure, censoring indicator and covariate vector one has data on the type of failure. As can be seen in Table 3 and Table 4 the systems under observation can fail due to availability, confidentiality and integrity related problems. The idea is now to consider cause-specific hazard functions, for each cause j there is an individual hazard function. The overall failure rate is defined as the sum over all cause-specific failure rates, i.e.

$$\lambda(t,Z)=\sum_{j=1}^{m}\lambda_j(t,Z) \tag{4.12}$$

The Kaplan-Meier estimates of the cause-specific failure rates of the three protected and the three unprotected systems are given in Figure 22. Note that the trends from the protected and unprotected systems cause-specific hazard models do not seem to differ too much from the one-covariate model. Visually it does not seem unlikely that the data obtained from the availability, confidentiality and integrity failures are from the same CDF.
Again a hypothesis test is proposed. This time it was chosen to test the hypothesis that the data obtained from the unprotected systems with the causes of failure availability, confidentiality and integrity are actually from the same survivor function. With the aid of Equations (4.6) and (4.7) the following failure vector is obtained

$$v=\begin{pmatrix}0.796957435\\4.791985695\\-5.58894313\end{pmatrix}$$

and variance matrix

$$V = \begin{pmatrix} 12.64 & -1.36 & -11.29 \\ -1.36 & 6.27 & -4.91 \\ -11.29 & -4.91 & 16.19 \end{pmatrix}$$

This yielded a value of 3.93 as approximation for the χ_2^2 statistic. The significance that was reported by Excel was 0.14. The null hypothesis that the availability, confidentiality and integrity failure data are from the same survivor function cannot be rejected (at the 0.05 level) on basis of this data as was already expected by consideration of the KM estimates for the survivor function.
In addition a log-rank test for the protected systems with the three different causes of failure is performed. The following failure vector was obtained

$$v = \begin{pmatrix} -1.60 \\ -3.32 \\ 4.92 \end{pmatrix}$$

with variance matrix

$$V = \begin{pmatrix} 10.42 & -2.60 & -7.82 \\ -2.60 & 9.67 & -7.07 \\ -7.82 & -7.07 & 14.90 \end{pmatrix}$$

With the aid of Equation (4.9) this yielded a value of 1.78 for the χ_2^2 statistic with significance 0.41. So the null hypothesis cannot be rejected (at the 0.05 level).

According to the log-rank test the possibility that the cause-specific failures are from the same CDF could not be rejected. Still it is not said that they are actually the same. Therefore I still propose the proportional hazard model since it could still be that the cause-specific hazards are proportional.

In this model with the cause specific hazard rates the partial likelihood can be written as

$$L\left(\hat{B}_1, \cdots \hat{B}_m\right) = \prod_{j=1}^{m} \prod_{d(t_{ji})>1} \frac{\exp\left(\hat{B}_j{}^T s_i\right)}{\left[\sum_{t_{ki} > t_{ji}} \exp\left(\hat{B}_j{}^T Z_k\right)\right]^{d(t_{ki})}} \tag{4.13}$$

Where t_{ji} is the i-th failure of cause j. It is stated that the coefficients can be calculated by conducting likelihood estimations for all m factors individually.
As in the one-covariate example, the Breslow estimator is taken in order to be able to deal with ties and the Newton-Raphson technique is used to estimate the coefficients $\hat{B}_i$. The estimates for the availability and integrity type failure were $\hat{B}_A$ = *-1.72* and $\hat{B}_I$ = *-1,74* after (again) 4 iteration with initial estimates 0. Coefficient $\hat{B}_C$ unfortunately could not be estimated because the protected systems did only failed after all the unprotected systems had

already failed, and, as a result, the maximum likelihood estimate is minus infinity. This is an awkward limitation of the partial-likelihood estimator and is in my opinion the result of too little data on this type of failure. An extra assumption that one can impose, such that the partial likelihood function has a maximum is that the failures should be mixed, i.e. for all Z_i there exists a t_i, such that $t_j < t_i < t_k$. Where $D_i = D_j = D_k$ and $Z_i \neq Z_j = Z_k$.

However, another strange anomaly was noted. In my opinion the covariates are strongly correlated, in fact they are the same! So, ideally, the estimates for each risk type are *–1.82*, as was the case in the one-covariate model. Or, if the causes really where different and contributed accordingly, one hopes for additivity of the estimator, in the sense that if one sums the coefficients from all types of failure one obtains the same value as one got in the estimate with only one covariate. Due to restrictions on the data, these ideas could not be tested.

Now that the cause-specific hazard rates were estimated with the assumption of proportionality, it is proposed to test if this assumption was valid. A plot of the cumulative hazards is supplied, see Figure 23.
In Figure 23 the cause-specific hazard rates are computed for all the KM estimators. Note that the hazard rates for the cause-specific failures of the protected and unprotected systems do not look proportional at all. From the cumulative hazard rates it is inferred that the level of protection of a system influences the survival of the systems. It seems that the protection of the systems has the same influence on the availability, confidentiality and integrity failure mechanisms.

4.6 Conclusions and recommendations from chapter 4

In this chapter the practical difficulties of handling reliability data were shown. To handle this kind of data the PHM was chosen. The crucial factor for choosing this model was the presence of a lot of censored observations as is typical in reliability data.
Although this model was able to illustrate the difference between protected and unprotected systems it was not able to capture the different types of the failures. This was on the one hand due to the form of the data, and on the other hand due to the choice of proportional hazards as opposed to time dependent hazards. However, in general it is difficult, if not impossible to rule out inference, and, the ability of the PHM to cope with inference of multiple pertinent covariates remains little. Nevertheless the model may still be well suited to indicate if the effects of covariates are positive or negative.
It was found that the Newton-Raphson method used to estimate the coefficients of the covariates does not always converge. If the failures accompanying a particular covariate are not mixed, as was described in section 4.5, the Newton-Raphson estimator diverges.

In general, however, it is however the opinion of the writer that inference makes interpretation of the coefficients of the covariates from such calculations difficult as they may be due to circumstantial causes. This is supported by the theoretical results from the next chapter.
As was suggested by multiple experts it might be that the problem as posed here is a competing risks problem and that this model will lead to better results in these kind of circumstances.

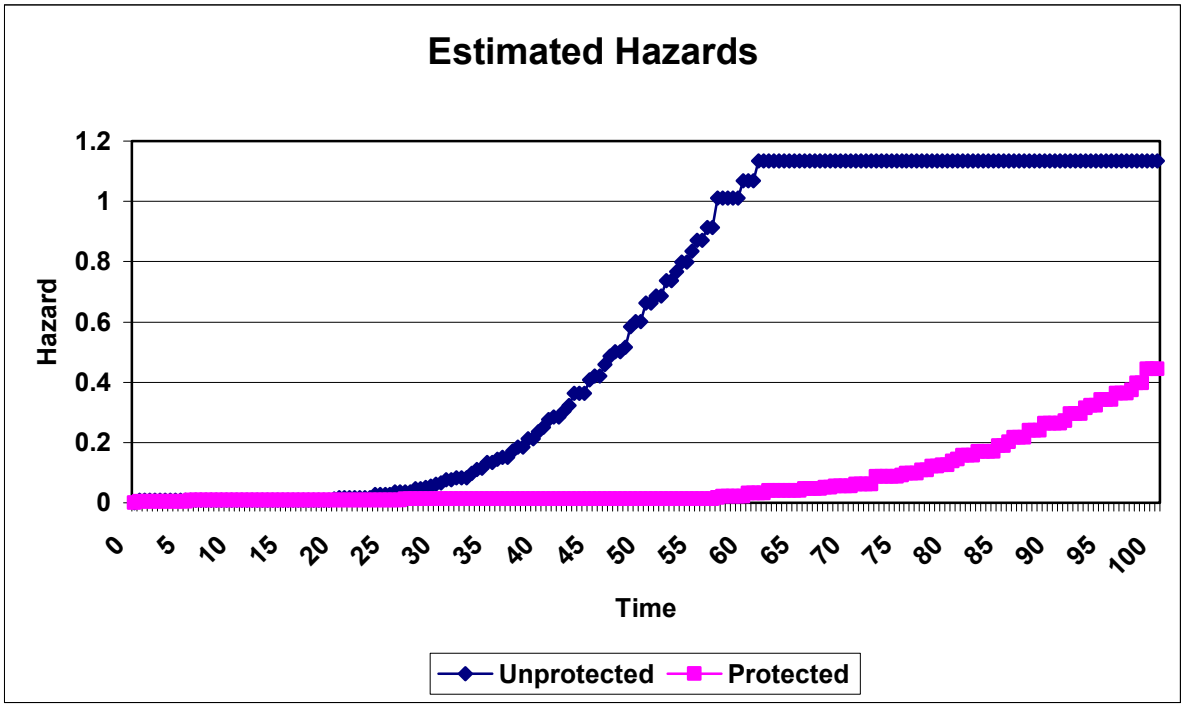

Figure 21: The estimated cumulative hazard curves obtained from the Kaplan-Meier survivor curves of the unprotected and protected system.

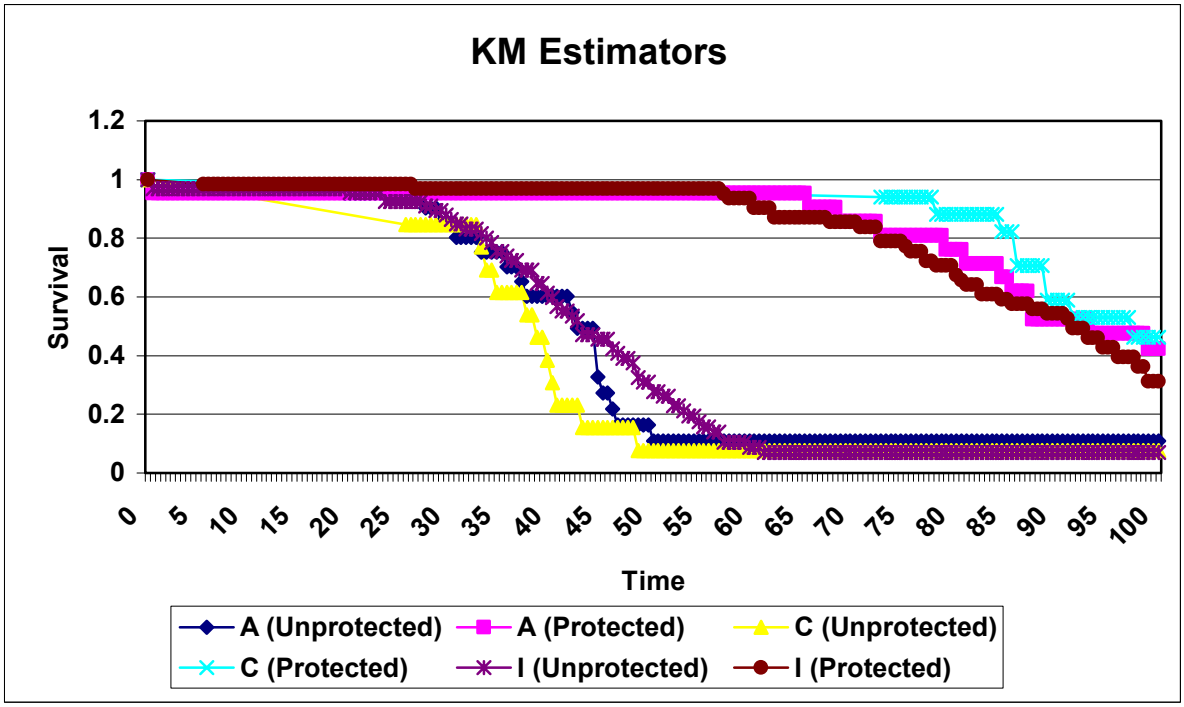

Figure 22: The KM estimates for the survival curves for the protected and unprotected cause-specific hazard rates.

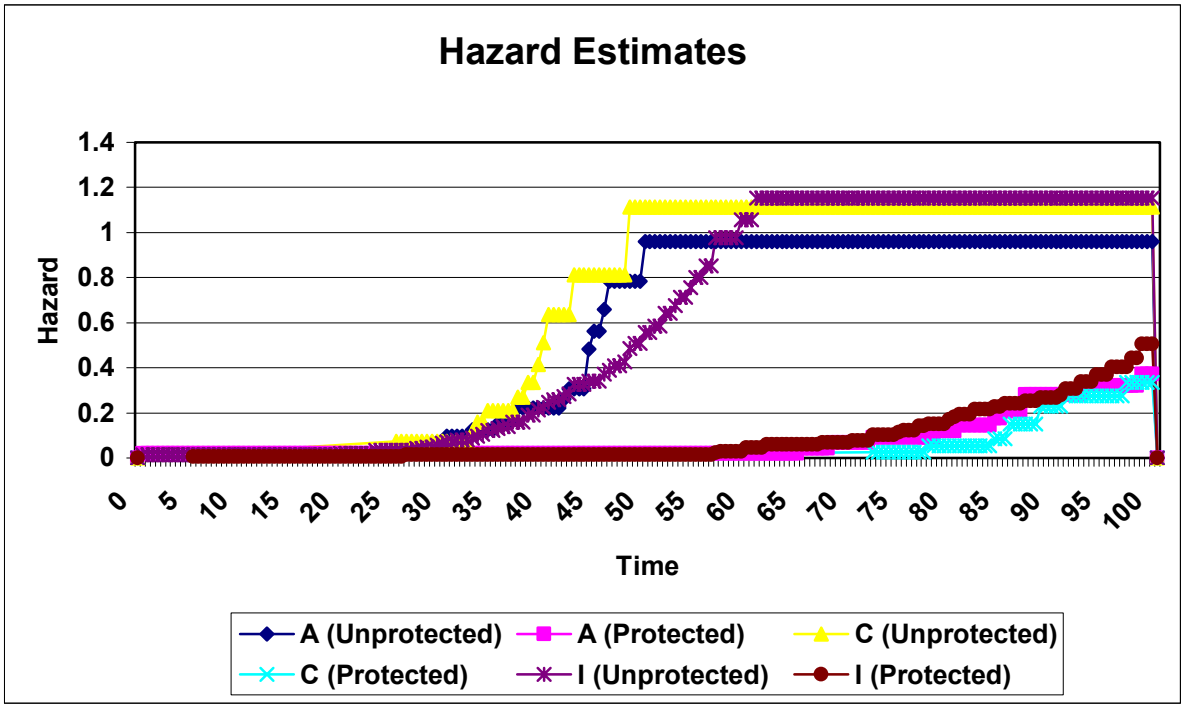

Figure 23: The estimated cumulative hazard curves obtained from the Kaplan-Meier survivor curves of the cause-specific hazard rates.

5 Missing Pertinent Dependent Covariates in the Proportional Hazard Model

This chapter elaborates on the estimation of parameters in the proportional hazard model. It suggests that Cox's models estimation of covariates may well be subject to bias implied by the number of covariates that are considered in the model.

To test the model adequacy of the proportional hazard model (PHM) multiple tests are available, see Cox [1972]. The Cox-Snell residual, Cox and Snell [1968], and Schoenfeld residual, Schoenfeld [1982], were introduced to test for the proportionality assumption and for possible outliers. Martingale residuals do basically the same but the martingale residual gives them a strong mathematical underpinning, see Therneau and Grambsch [1990]. For a good overview of the different residuals using life data, see for instance Ansell and Phillips [1997]. The PHM seems to induce bias on itself. In most regression models with independent explanatory variables, one can drop some while the estimates of the others stay consistent as for instance linear regression, for a good overview on linear regression, see Miles and Shevlin [2001].

However in Cox's PHM even omitting an independent, pertinent covariate biases the estimates of the other covariates, as was also found by Bretagnolle and Huber [1988]. As a result the estimates are shown to have a bias toward the null in the case of independence. The bias resulting from omitting covariates is not negligible and may even result in confidence intervals that do not contain the actual value, see Cooke and Morales [2004]. A common recommendation is that the inclusion of additional covariates reduces the bias.
Gerds [2001] showed that in specific cases correlation between covariates can even induce more bias. Moreover it was shown that functional misspecification can also induce a bias. A possible solution was proposed by Abrahamowicz [2004], who had simulation results that showed that covariate aggregation can significantly reduce the bias as compared to covariate emission.
In this chapter it was found that for negatively correlated covariates, omitting one of the covariates results in a bias. It was proven that the bias is towards the null.

In section 5.1 the framework for this chapter is given. In section 5.2 the results from Bretagnolle are repeated more strictly. It shows that when one omits relevant variables it is possible to compute the sign of the bias for two different cases. Section 5.3 elaborate on the results from section 5.2 and considers dependent covariates. It will be proven that in the case of negative correlated covariates, omitting a single one also results in a bias toward the null. Finally, section 5.4 gives brief conclusions and recommendations.

5.1 Model framework

In this chapter the framework is given upon which the results of this chapter are based. In subsection 5.1.1 an general overview of the PHM is given and in subsection 5.1.2 the two different models that will be compared are defined.

5.1.1 General Cox Proportional Hazard Model

Cox's proportional hazard model is a regression model for predicting survival time X, where X is a positive random variable denoting the time from which an individual is studied until termination of the study, failure or censoring from the variable. The appeal of the model is the use of covariates Z, which is denoted in vector form and assume given by $(Z_1,..., Z_k)^T$. The distribution function of which is H.
Letting h stand for the hazard rate, the formal definition of which is

$$h(t) = \lim_{\Delta t \downarrow 0} \frac{\Pr\{t \le X < t + \Delta t \,|\, X \ge t\}}{\Delta t}.$$

In the proportional hazard model the hazard function is assumed to be of the following form

$$h(t, Z) = h_0(t) \exp\{B^T Z\} \tag{5.1}$$

where h_0 is the *basic hazard rate*. It is noted that the basic hazard rate can be interpreted as the risk an individual with all covariates zero runs.
The parameter B is the vector of coefficients of the covariates, the k-dimensional parameter that is of primary interest in Cox's model. Note that positive values for the coefficients indicate unfavorable covariates and negative values favorable. For all coefficients non-zero the corresponding covariate is called *pertinent*.
Given the hazard rate the survival function S of the individual, could be calculated and is defined by the exponent of minus the integral of hazard function from zero up to time t.

$$S(t, Z) = \exp\left\{-\int_0^t h(u, Z)\, du\right\} \tag{5.2}$$

And since the survival function is defined by Pr(X>t), the distribution function is also known.
An interesting detail of the model is that the hazard can be expressed as minus the fraction of the survival functions derivative over the survival function

$$h = -\frac{S'}{S}$$

So, given h, S or F the other two are also known.
In the model the observed period is subject to right censoring by positive random variable C with survival function G, this random variable is considered independent of X conditional on Z. This leads to the formula for the time, T that is observed

$$T = \inf(X, C). \tag{5.3}$$

The data that is obtained during an observation is typically of the form $(T,D,Z)_i$ for $i=1,...,n$. Where D is an indicator function taking value $D = 1$ if the individual dies and $D = 0$ if the individual is censored.

In order to estimate the coefficients of the covariates B, Cox proposes an ingenious estimate the partial likelihood that is a product of conditional likelihoods

$$L(B) = \prod_{i=1}^{n} \left(\frac{e^{B^T Z^i}}{\sum_{j \geq i}^{n} e^{B^T Z^j}} \right)^{D_i} \tag{5.4}$$

In this way Cox gets rid of the basic hazard rate parameters, which in the maximization of a marginal likelihood approach have to be considered, see Kalbfleish and Prentice [2002] and Lawless [1987]. The estimator is proven to be consistent and asymptotically normal.

5.1.2 Complete and incomplete model

The effect of missing pertinent covariates in the model is investigated. In usual linear regression some explanatory variables could be dropped, which does not affect the estimates of the other variables. In Cox's Proportional Hazard Model however this is not the case. Under specific assumptions the sign of the bias can even be predicted.

In the following two models will be compared, the first of which is a complete Cox's PHM, i.e. a model with all the pertinent explanatory variables and the second is a model not containing all pertinent covariates. The following notation is adopted for the covariate vector of the *complete model*

$$Z = \left(Z' Z^* \right)^T = \left(Z'_1, \cdots, Z'_{k'}, Z^*_1, \cdots, Z^*_{k*} \right)^T . \tag{5.5}$$

The covariate vector is a k-dimensional vector, with $k = k' + k^*$. And for the vector of the coefficient of the covariates similar notation is used, that is to say $B = (B'B^*)^T$. The covariate vector and the coefficient vector of the *incomplete model* are the first parts of the particular vectors of the complete model, that is to say Z' and B' respectively. In short the complete model for the hazard rate is given by

$$h(t, Z) = h_0(t) \exp\left\{ B^T Z \right\} = h_0(t) \exp\left\{ \left(B' \right)^T Z' + \left(B^* \right)^T Z^* \right\}. \tag{5.6}$$

And the hazard rate for the incomplete model is of the form

$$h(t, Z) = h_0(t) \exp\left\{ \left(B' \right)^T Z' \right\}. \tag{5.7}$$

It is noted that the complete model can be expressed as a product of the incomplete model and an exponent of new covariates.

5.2 Independent Covariates

In this part the covariates are considered independent. It is shown that if the number of pertinent covariates is miss-estimated the estimation of the covariates themselves is not consistent.
In subsection 5.2.1 an ingenious method is proposed for calculating the sign of the bias and in subsection 5.2.2 it is shown that for two different cases the method proposed in subsection 5.2.1 applies. In subsection 5.2.3 an upper time-bound is supplied for which the results of the previous paragraph hold. Finally, in subsection 5.2.4 the assumptions are stated, which need to be satisfied when the results of the previous paragraphs are supposed to hold.

5.2.1 The sign of the bias

In Bretagnolle it was showed that when Fischer's information matrix has additional properties, the sign of the bias of the estimates of the covariate can be predicted. This result is stated in the proposition given below. First some notational aspects are clarified.

Let $b = (b'b^*)$ be the vector of true values for the covariates. Furthermore let $B'(b^*)$ denote the value that maximizes the likelihood for the complete model and $B'(0)$ is the value that maximizes the likelihood for the incomplete model.
Furthermore, write $B^* = yb^*$ for all y taken from the interval $[0,1]$, so that for $y = 0$ this corresponds with the incomplete model and for $y = 1$ with the complete model.
Having clarified the notational aspects the proposition can be given.

Proposition 5.2.1 If Fischer's information matrix is of (DP+) type, and all components of b^* are strictly positive, then every component of the bias

$$B'(0) - B'(b^*) = -\int_0^1 \frac{dB}{dy}(yb^*)\,dy$$

is strictly negative.

Proof: By assumption $B'(yb^*)$ solves the following set of equations

$$\frac{\partial}{\partial B_i} L(B'(yb^*), yb^*) = 0 \tag{5.8}$$

for i=1,…,k'. Let the information be given and adopt the following notation

$$I = \begin{pmatrix} M' & N \\ N^T & M^* \end{pmatrix} \tag{5.9}$$

where M' corresponds to the k' by k' information matrix of covariates corresponding with the incomplete model and M^* the k^* by k^* sub matrix of omitted covariates. Let $M'(y)$ and $N(y)$ denote the matrices M' and N' evaluated at $B = (B'(yb^*), yb^*)$, respecitively.

Taking the derivative with respect to y of the equation (5.9) results in the following equation

$$\frac{dB'}{dy} = -\left[M(y)\right]^{-1} N(y) b^* \tag{5.10}$$

So if the information matrix is of (DP+) type at every value for B this also holds for $M'(y)$. And a property of this matrix is that all terms of its inverse are strictly positive, which is proven in lemma 5.2.3. Furthermore since $N(y)$ is out of the diagonal all its terms are negative and by assumption b^* has all its terms positive, so that the term is positive for all y taken from the interval $[0, 1]$. Combining this result with the formula in the proposition gives the result.

QED

An assumption in proposition 5.2.1 was that Fischer's information matrix has to be of (DP+) type. In proposition 5.2.7 it is shown that when the model satisfies the assumptions listed in subsection 5.2.4 Fischer's information matrix always will be of this type.
The definition for a (DP+) type of matrix is given below.

Definition 5.2.2 A square matrix M is called of (DP+) type if it is symmetric, positive definite and if its general term $M_{i,j}$ is strictly negative for all $i \neq j$. It is noted that such a matrix has a positive main diagonal.

One of the reasons the matrix should be of this type is that the inverse of this type of matrix is always strictly positive, as is stated in the following lemma.

Lemma 5.2.3 If M is of (DP+) type then each term of its inverse is strictly positive.

Proof: Choose diagonal matrix D in the following way.

$$D = \begin{pmatrix} M_{11} & 0 & \cdots & 0 \\ 0 & M_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & M_{nn} \end{pmatrix}^{-1/2} \tag{5.11}$$

Then the product of D and M can be written in the following way

$$DMD = I - A \tag{5.12}$$

where A is a matrix with zeros on the diagonal and strictly positive terms on the other entries. Furthermore I - A is symmetric positive definite. Since D is positive definite too and the product of positive definite matrices stays positive definite. Recall the theorem from classical analysis that

$$\left(I - A\right)^{-1} = \lim_{n\to\infty} \sum_{n} A^n \tag{5.13}$$

So that the inverse is strictly positive since every term in A is positive, making sure the sum is also if the limit exists. Since proper parameterization is assumed, see assumption (H1), matrix

A has full rank and is thus diagonizable and can thus be written as a product of a matrix with all eigenvectors P, a diagonal matrix with all eigenvalues on it D_A and the inverse of the eigenvector matrix. Thus the above simplifies into

$$\sum_n A^n = P\left(\sum_n D_A{}^n\right)P^{-1} \tag{5.14}$$

The sum above converges if every absolute value of the eigenvalues of matrix A is smaller than one. Note that the $I - A$ is positive definite, so that for every non-zero vector V, $V^TAV<V^TV$, so the eigenvalues of A are smaller than 1. Since A is real and all entries off diagonal are positive, the matrix is irreducible. According to the Perron Frobenius theorem matrix A has a positive eigenvalue and furthermore it dominates all absolute values of each other eigenvalue, making sure the sum converges.
The inverse of matrix M can now be written as

$$M^{-1} = D\left(I - A\right)^{-1} D = D\left[\sum_n A^n\right]D \tag{5.15}$$

And since this is a product of positive matrices every entry of the inverse matrix of M is shown to be positive, as was required.

QED

In the proof of lemma 5.2.3 the Perron Frobenius theorem was used. For completeness a simplified version of the Perron Frobenius theorem is given below. But first another definition is given, which defines the reducibility used in the theorem. A definition of irreducible matrices is given below

Definition 5.2.4 Let A = (a_{ij}) be from $\mathbb{C}_{n\times n}$ and let Z be the set of the first n positive integers

$$Z = \{1, 2, \cdots, n\}.$$

A is reducible it there exists a partition of the set Z into two nonempty disjoint subsets S and T such that $a_{ij} = 0$ for all $i \in S, j \in T$. A matrix that is not reducible is irreducible.

So, now the Perron Frobenius theorem can be understood fully.

Perron Frobenius Theorem If A = (a_{ij}) is an irreducible nonnegative matrix, then A has a positive real eigenvalue λ_+ with the following properties

1. For all other eigenvalues λ of the matrix A the following holds $|\lambda| < \lambda_+$.
2. The eigenvector belonging to the eigenvalue λ_+ can be chosen to have positive components.

Proof: Proofs can be found in Gantmacher (1960) or Varga (1962).

It is noted that the definition above is not useful in showing that matrices are reducible or not. In order to show that general matrices are irreducible the following theorem helps.

Theorem 5.2.5 Let A=(a_{ij}) be from $\mathbb{C}_{n\times n}$. The matrix is said to be irreducible if and only if its associated directed graph *G(A)* is strongly connected.

Proof: A proof can be found in Usmani 1987.

And for completeness the definition of a directed graph is also provided.

Definition 5.2.6 A directed graph is strongly connected if for any ordered pair of nodes P_i and P_j there exists a path that connects P_i to P_j.

Note that this definition is of much help in noticing that a matrix is irreducible or not. Because if all the entries in the matrix are unequal to zero the matrix is sure to be irreducible.

5.2.2 Form of the information matrix

Essentially the proof of the sign of the bias hinges on the information matrix being of (DP+) type, which is shown in the proposition below. First however, some notational remarks are given.

Since Cox's partial likelihood does not depend on time, furthermore, put $dt = h_0(u)du$, which leaves the likelihood unchanged, but gets rid of the basic hazard rate. For convenience put $G = G(t,z)$ and $S = S(t,z)$. With the aid of the law for large numbers the limiting log partial likelihood is given by

$$L(\beta) = \int E_Z \left[\left[B^T Z - \log E_Z \left\{ \exp\left(B^T Z\right) GS \right\} \right] MGS \right] \mathrm{dt} \tag{5.16}$$

where

$$M = \exp\left(b^T Z\right) \tag{5.17}$$

the true effect of the covariate *Z*. Note that although the time was changed it was chosen to still denote the survivor function of the censoring *G*, for convenience sake.
Let D_i denote the derivative to B_i. The first derivative of the limiting log partial likelihood is given by

$$D_i L(B) = \int E_Z \left[\left(Z_i - E_Z \left[Z_i U \right] / E_Z \left[U \right] \right) MGS \right] dt, \tag{5.18}$$

where

$$U = \exp\left(B^T Z\right) G(t,Z) S(t,Z). \tag{5.19}$$

And the second derivative with respect to covariates B_i and B_j is given by

$$D_{ij}^2 L(\beta) = -\int E_Z\left[\left(E_Z\left[Z_i Z_j U\right]E_Z[U] - E_Z[Z_i U]E_Z\left[Z_j U\right]\right)E_Z[U]^{-2}\, MGS\right]dt \quad (5.20)$$

the product of the estimate of the time independent part of the hazard rate, the survivor functions of the censoring and the survivor function of the survival time.
As was said above, in case that the Fischer information matrix is of (DP+) type, the sign of the bias can be predicted. This is done in proposition 5.2.7 below. A list of the assumptions used for this proposition can be found in section 5.2.4.

Proposition 5.2.7 Form of the information matrix.

a. If is assumed that the assumptions H0, H1, H2 and H3' hold. Then at every value of the coefficients of the covariates the information matrix is positive definite. Thus the log partial likelihood function is strictly concave.
b. Assume that H0, H1, H2, H3' and H4' hold, and that b_i is positive for all i.
b1. If k = 2, then at every value the information matrix is of (DP+) type.
b2. For k > 2 and Z bounded. And let t_0 be a fixed constant. Then the same property for the information matrix holds as long as $P[C<t_0] = 1$.
c. Under the same assumptions as b. If $b_i \neq 0$ for all i, then $\text{sign}(-D^2_{ij}) = -\text{sign}(b_i)\text{sign}(b_j)$ for $i \neq j$. If for some i, $b_i = 0$, then $D^2_{ij}=0$ for $i \neq j$. But the diagonal elements of the information matrix are always positive.

Because of the mathematical nature of the proof it is supplied in the appendix. The proof above uses the theorem of Cauchy-Schwarz for completeness the used version of this theorem is provided below.

Theorem Cauchy-Schwarz Let f and g be two square-integrable functions in Euclidian space. Then the following holds

$$\left[\int |f(x)g(x)|dx\right]^2 \leq \int |f(x)|^2 dx \int |g(x)|^2 dx.$$

It is noted that if the right-hand-side dominates the absolute value of the left-hand-side and as a direct result it also dominates the left-hand-side without taking the absolute value.

In part b2 from proposition 5.2.7 a constant t_0 was used. It is noted that it is possible to compute the upper bound for the time; this is done in the next subsection.

5.2.3 Upper bound for the time

In this subsection the upper bound for the time t_0 is calculated. Substituting the expressions for D and D'' in equation (6.41) leads to

$$t \leq \left(\exp\left(B^T sZ\right) + \exp\left(B^T sZ''\right)\right)\left(\exp\left(B^T sZ\right) - \exp\left(B^T sZ''\right)\right)^{-2}. \quad (5.21)$$

Rearranging the terms on the right-hand-side results in

$$t \le \frac{\left(\frac{\exp\left(B^T sZ\right)}{\exp\left(B^T sZ''\right)}+1\right)}{\left(\frac{\exp\left(B^T sZ\right)}{\exp\left(B^T sZ''\right)}-1\right)^2 \exp\left(B^T sZ''\right)}. \tag{5.22}$$

Note that this has to hold for every possible *s*-transformation and every possible value of *Z*. Thus by calculating the lower bound of the right-hand-side the upper bound t_0 for the time is obtained

$$t_0 = \left(e_m + 1\right)\left[\left(e_m - 1\right)^2 \exp\left(ess\sup\left\{b^t Z\right\}\right)\right]^{-1} \tag{5.23}$$

where

$$e_m^{-1} = \exp\left(ess\sup\left\{b^t Z\right\} - ess\inf\left\{b^t Z\right\}\right) \tag{5.24}$$

Remind that the essential supremum of a function *f* is the number *a*, such that *f* exceeds *a* only on a set of measure zero and the essential infimum is defined in a similar way. Note that e_m is the minimum effect of the covariates, which is bounded since it was assumed that *Z* was bounded. And that t_0 is the upper bound for *t* derived in proposition 2.

For completeness the definition of the essential supremum used in equation (5.24) is given.

Definition Given a measurable function $f : X \to \Re$, where *X* is a measure space with measure *P*, the essential suppremum is the smallest number *a* such that the set *{f(x):f(x)>a}* has measure zero. If no such number exists, choose for instance *f(x)=1/x*, then the essential supremum is infinity.

It is noted that the essential supremum is a generalization of the maximum. The difference is that the values the function takes on a set of measure zero do not affect the essential supremum.
The essential infimum is defined in a similar way as the essential supremum, but in this case giving a lower bound for the function except for again a set of measure zero.

5.2.4 Assumptions

In the proofs of the propositions various assumptions where made. In this part the assumptions are listed to get an overview. The five initial assumptions are listed below.

(H0) The covariate vector Z is time independent, with distribution function H and finite exponential moments. In addition *C* is independent of *X* given *Z*.
(H1) Proper parameterization: there is no proper linear subspace including *Z* almost surely.
(H2) From all individuals in the study with equal covariates, you see at least one dieing, that is to say Pr(C<X|Z) <1 almost surely.
(H3) All the components of the studied vector *Z'* are independent and is moreover independent of the omitted components of Z^*.

(H4) The survivor function of the censoring can be written as a product.

$$G(t,Z) = \Pr\{C \geq t | Z\} = \left[\prod_{i=1}^{k'} G_i(t,Z_i)\right] G^*(t,Z^*) \tag{5.25}$$

(H5) Mixture of failures, i.e. for all Z_i there exists an t_i, such that $t_j < t_i < t_k$. Where $D_i = D_j = D_k$ and $Z_i \neq Z_j = Z_k$.

In the part below an interpretation for the assumptions is given.
(H0) is assumed to hold, because it sets the boundaries of the results in this paper and the assumptions are not so rigorous that they seem stringent.
Also (H1) is assumed to hold since otherwise Fischer's information matrix does not have full rank, so that the matrix is singular and the inverse does not have to exist.
Hypothesis (H2) seems unexpected, since one could think that it is sufficient for the censoring and the survival time being independent. But if you have a set of individuals with the same particular covariates and all of the individuals are censored, then it is difficult to make a guess for the value of the covariate. Which results in Cox's partial likelihood being zero or infinity, depending on the positive or negative effect of the covariate, and one has to come up with some other estimator.
The hypothesis (H3) is a bit restrictive, but is necessary for the calculations to hold.
Hypothesis (H4) is some technical assumption that is needed further along in proof of proposition 5.2.7.
Hypothesis (H5) makes sure that the likelihood function is concave.

For some of the proofs there sharpened versions of the assumption are needed. Furthermore, define the sharpened versions of (H3) and (H4)
(H3') All components of Z are independent
(H4') The entire survivor function of the censoring can be written as a product

$$G(t,Z) = \Pr\{C \geq t | Z\} = \prod_{i=1}^{k} G_i(t,Z_i) \tag{5.26}$$

5.3 Dependent covariates

In this chapter the results from the previous chapter are elaborated on. The general idea is the same as in the previous chapter. Again a complete and an incomplete model are compared, but in this case the covariates may be dependent.

5.3.1 Formula for the bias

The idea is roughly similar to the idea stated in paragraph 4.1. As a direct result from proposition 1 it follows that it is possible to calculate the bias of the incomplete model. However the results of proposition 1 only hold in special cases. Recall that the information matrix of the complete model can be written in the following form

$$I = \begin{pmatrix} M' & N \\ N^T & M^* \end{pmatrix}. \tag{5.27}$$

An assumption is that Fischer's information matrix should be of (DP+) type, which is to my reckoning a stringent one. So, dropping the assumption that the information matrix is of (DP+) type and also dropping the assumption on *b* leads to the following proposition.

Proposition 3 The bias of the incomplete model is of the form

$$B'(0) - B'(b^*) = -\int_0^1 \frac{dB}{dy}(yb^*)\,dy, \tag{5.28}$$

where the derivative is a product of parts of the complete information matrix and the omitted covariates

$$\frac{dB'}{dy} = -\left[M'(y)\right]^{-1} N(y) b^*. \tag{5.29}$$

Proof: This is a direct result of proposition 1.

5.3.2 The Two-dimensional case

Assume to have a complete model of dimension two with dependent covariates. It is shown that when the correlation is negative Fischer's information matrix still is of (DP+) type. Define the correlation between the two covariates as given by ρ, that is to say

$$\frac{dZ_1}{dZ_2} = \rho. \tag{5.30}$$

Recall that according to Equation (5.16) the elements of the information matrix can be written as

$$D_{ij}^2 L(\beta) = -\int E_Z\left[\left(E_Z\left[Z_i Z_j U\right] E_Z[U] - E_Z\left[Z_i U\right] E_Z\left[Z_j U\right]\right) K\right] dt \tag{5.31}$$

where

$$K = E_Z[U]^{-2}\, MGS. \tag{5.32}$$

With *M* according to Equation (5.17), *G* the survivor function of the censoring and *S* the survivor function of the individual.
The case when $i = j = 1$ is considered. As a result, Equation (5.31) transforms into

$$D_{11}^2 L(\beta) = -\int E_Z\left[\left(E_Z\left[Z_1^2 U\right] E_Z[U] - E_Z\left[Z_1 U\right]^2\right) E_Z[U]^{-2}\, MGS\right] dt. \tag{5.33}$$

Define Z_3 as

$$Z_3 = Z_1 - \rho_{12} Z_2. \tag{5.34}$$

Such that Z_3 is uncorrelated with Z_2. Substitution of Equation (5.34) in Equation (5.33) results in

$$D_{11}^2 L(\beta) = -\int E_Z \left[\left(E_Z \left[\left(Z_3^2 + 2\rho Z_3 Z_2 + \rho^2 Z_2^2 \right) U \right] E_Z [U] \right) E_Z [U]^{-2} \, MGS \right] dt$$
$$+ \int E_Z \left(E_Z \left[(Z_3 + \rho Z_2) U \right] E_Z \left[(Z_3 + \rho Z_2) U \right] \right) dt$$

Rearranging terms leads to the following expression

$$D_{11}^2 L(\beta) = A_1 + A_2 + A_3$$

where

$$A_1 = -\int E_Z \left[\left(E_Z \left[\left(Z_3^2 \right) U \right] E_Z [U] - E_Z [Z_3 U]^2 \right) E_Z [U]^{-2} \, MGS \right] dt$$
$$A_2 = -\rho^2 \int E_Z \left[\left(E_Z \left[\left(Z_2^2 \right) U \right] E_Z [U] - E_Z [Z_2 U]^2 \right) E_Z [U]^{-2} \, MGS \right] dt$$
$$A_3 = -2\rho \int E_Z \left[\left(E_Z \left[(Z_3 Z_2) U \right] E_Z [U] - E_Z [Z_3 U] E_Z [Z_2 U] \right) E_Z [U]^{-2} \, MGS \right] dt$$

It is claimed that A_1 and A_2 are always positive. Because Z_2 and Z_3 are independent the corresponding information matrix of these two covariates would have diagonal elements A_1 and A_2 / ρ^2, respectively, and these diagonal elements would be positive as was shown in Proposition 5.2.7.
In a similar way it can be shown that A_3 is positive provided that $\rho < 0$. As the off diagonal term of the information matrix corresponding to covariates Z_2 and Z_3 would have the element $A_3 / (2 \rho)$, which was shown to be negative. So that the expression in Equation (5.33) is always positive.
Note that in a similar fashion it can be shown that D^2_{22} is always positive. So that the diagonal of a two dimensional Fisher matrix is shown to be positive, as was required.
Remains to show that the off-diagonal terms are negative. Consider the off-diagonal term

$$D_{12}^2 L(\beta) = -\int E_Z \left[\left(E_Z \left[(Z_1 Z_2) U \right] E_Z [U] - E_Z [Z_1 U] E_Z [Z_2 U] \right) E_Z [U]^{-2} \, MGS \right] dt \tag{5.35}$$

Again substituting Equation (5.34), and rearranging terms of this expression yields

$$D_{12}^2 L(\beta) = C_1 + C_2$$

where

$$C_1 = -\rho \int E_Z \left[\left(E_Z \left[\left(Z_2^2 \right) U \right] E_Z [U] - E_Z [Z_2 U]^2 \right) E_Z [U]^{-2} \, MGS \right] dt$$
$$C_2 = -\int E_Z \left[\left(E_Z \left[(Z_3 Z_2) U \right] E_Z [U] - E_Z [Z_3 U] E_Z [Z_2 U] \right) E_Z [U]^{-2} \, MGS \right] dt$$

Note that C_2 is always negative, since the off-diagonal term of the information matrix corresponding to covariates Z_2 and Z_3 would have this element.
Moreover, for $\rho < 0$, C_1 will also be negative as the same information matrix would have the positive element C_1 / ρ. Such that as a result the expression in Equation (5.35) is always negative.
Note that the other off-diagonal term of the information matrix has the same term, such that the Fischer information matrix of the model with two negatively correlated covariates is of (DP+) type.

5.4 Conclusions and recommendations from chapter 5

The PHM finds wide use in assessing covariates in reliability data. There exists numerous tests to assure model accuracy and pertinence of covariates. However, the PHM seems to be unable to cope with inference. And as a result, direct interpretation of coefficients of the covariates remains controversial.
In this chapter it was shown that omission of pertinent covariates in case of negative correlation biases the estimates of the covariates toward the null as was already known for the case of independent covariates. Furthermore closed formulations have been derived for the entries in Fischer's information matrix in case of dependent covariates.
All in all the proportional hazard model seems subject to bias which is unpredictable if not all concomitant variables are known. This seems to indicate a major drawback of the model. Further research is recommended to stretch results to more dimensions, or even expressions for positively correlated variables.

In general estimation of covariates is difficult if correlation exists. It is the authors opinion that assumptions of independence need to be backed by evidence or at least have a solid theoretical foundation. In other cases it is my perception that covariate values may well be misinterpreted due to the ignorance regarding the dependency of variables.

6 Appendix

The idea's for the various proofs of the Poisson processes come from Wikipedia [2005] and Nenadic [2005].

Proof of Equation (2.2)

Let variable X follow a binomial distribution with parameters n and $\lambda t/n$, then

$$\lim_{n\to\infty} Pr\left(X=k\right) = \lim_{n\to\infty}\binom{n}{k} p^k \left(1-p\right)^{n-k} = \lim_{n\to\infty} \frac{n!}{k!\left(n-k\right)!}\left(\frac{\lambda t}{n}\right)^k \left(1-\frac{\lambda t}{n}\right)^{n-k}. \tag{6.1}$$

Rearranging terms gives

$$\lim_{n\to\infty} \underbrace{\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-k+1}{n}\right)}_{A} \underbrace{\frac{\left(\lambda t\right)^k}{k!}}_{B} \underbrace{\left(1-\frac{\lambda t}{n}\right)^n}_{C} \underbrace{\left(1-\frac{\lambda t}{n}\right)^{-k}}_{D}. \tag{6.2}$$

Now note that both A and D go to 1 as n goes to infinity and that B is constant in n. So that by some elementary calculus the following result is obtained

$$\lim_{n\to\infty} Pr\left(X=k\right) = \frac{\left(\lambda t\right)^k}{k!} e^{-\lambda t}. \tag{6.3}$$

QED.

Proof of theorem 2.3.2

In this proof it is shown that the moment generating functions of the homogeneous Poisson process equals that of the Poisson distribution and that hence by uniqueness the proof is concluded.

Recall that the moment generating function of a random variable X is given by

$$M_X\left(z\right) = E\left(e^{zX}\right), \qquad \mathrm{z} \in \mathbb{R}, \tag{6.4}$$

wherever this expectation exists.
Let Y be Poisson distributed with parameter λ, then with the aid of Equation (6.4) the moment generating function can be written as

$$M_Y\left(z\right) = E\left(e^{zY}\right) = \sum_{i=0}^{\infty} E\left(e^{zi} p_i\right)$$

where $p_i = P(X=i)$. And since Y is a discrete random variable this results in

$$\sum_{i=0}^{\infty} e^{zi} \frac{(\lambda t)^i}{i!} e^{-\lambda t} = e^{\lambda t\left(e^z-1\right)} \tag{6.5}$$

The moment generating function of the Poisson process at $t+\Delta t$ can by independence be written as

$$M_{N(t+\Delta t)}(z) = E\left[e^{zN(t+\Delta t)}\right] = E\left[e^{zN(t)+zN(t+\Delta t)-zN(t)}\right] = E\left[e^{zN(t)}\right] E\left[e^{z\left(N(t+\Delta t)-zN(t)\right)}\right].$$

With aid of the definition of the Poisson process this can be simplified into

$$M_{N(t)}(z)\, E\left[e^{zi} p_i\right] = M_{N(t)}(z)\left(1-\lambda\Delta t - o(\Delta t) + e^z(\lambda\Delta t) + e^{2z} o(\Delta t) + e^{3z} o(\Delta t) + \cdots\right). \tag{6.6}$$

Consider the following formula

$$\frac{M_{N(t+\Delta t)}(z) - M_{N(t)}(z)}{\Delta t}. \tag{6.7}$$

Substituting this result obtained in Equation (6.6) leads to

$$\begin{aligned}
&\frac{M_{N(t+\Delta t)}(z) - M_{N(t)}(z)}{\Delta t} = M_{N(t)}(z)\left(-\lambda - \frac{o(\Delta t)}{\Delta t} + e^z\left(\lambda\frac{\Delta t}{\Delta t}\right) + e^{2z}\frac{o(\Delta t)}{\Delta t} + e^{3z}\frac{o(\Delta t)}{\Delta t} + \cdots\right) \\
&\Rightarrow \lim_{\Delta t\to 0} \frac{M_{N(t+\Delta t)}(z) - M_{N(t)}(z)}{\Delta t} = M_{N(t)}(z)\,\lambda\left(e^z - 1\right) \\
&\Rightarrow \frac{dM_{N(t)}(z)}{dt} = M_{N(t)}(z)\,\lambda\left(e^z - 1\right) \\
&\Rightarrow \log\left(M_{N(t)}(z)\right) - \underbrace{\log\left(M_{N(0)}(z)\right)}_{=0} = \lambda t\left(e^z - 1\right) \\
&\Rightarrow M_{N(t)}(z) = e^{\lambda t\left(e^z-1\right)}.
\end{aligned} \tag{6.8}$$

QED.

Proof of corollary 2.3.3

Take a Poisson process with parameter $\lambda > 0$, then the variance equals the first derivative of the moment generating function with respect to z evaluated at $z = 0$.

$$\left[\frac{dM_{N(t)}(z)}{dz}\right]_{z=0} = \left[\frac{de^{\lambda\left(e^z-1\right)}}{dz}\right]_{z=0} = \left[\lambda e^z e^{\lambda\left(e^z-1\right)}\right]_{z=0} = \lambda \tag{6.9}$$

This concludes the first part of this proof.

For the second part remember that the variance is defined as the difference between the second moment of the random variable and the square of the expectation, that is to say

$$Var(X) = E(X^2) - E(X)^2. \tag{6.10}$$

Thus for the calculation of the variance the second derivative with respect to z of the moment generating function evaluated in 0 is needed, this is given by

$$\begin{aligned} &\left[\frac{d^2 M_{N(t)}(z)}{dz^2}\right]_{z=0} = \left[\frac{d}{dz}\lambda e^z e^{\lambda(e^z-1)}\right]_{z=0} \\ &= \left[\lambda e^z e^{\lambda(e^z-1)} + \lambda^2 e^{2z} e^{\lambda(e^z-1)}\right]_{z=0} = \lambda + \lambda^2. \end{aligned} \tag{6.11}$$

So that by substitution of Equation (6.9) and Equation (6.11) in Equation (6.10) the proof is completed.

QED

Proof of theorem 2.3.4

Consider two independent homogeneous Poisson processes, X_1 and X_2, with intensities λ_1 and λ_2, respectively. By writing out the probability distribution for the superimposed process it is shown that the superimposed process, Y, is also a homogeneous Poisson process with intensity $\lambda_1 + \lambda_2$.
Writing out the probability that k events occur in time interval $[0,t)$ leads to

$$P(Y=k) = \sum_{i=0}^{k} P(X_1 = i, X_2 = k-i) = \sum_{i=0}^{k} P(X_1 = i) P(X_2 = k-i). \tag{6.12}$$

by conditioning on X_1 and independence. Substituting the probabilities from theorem 2.3.1 leads to

$$\begin{aligned} &\sum_{i=0}^{k} \frac{(\lambda_1 t)^i}{i!} e^{-\lambda_1 t} \frac{(\lambda_2 t)^{k-i}}{(k-i)!} e^{-\lambda_2 t} = \sum_{i=0}^{k} \frac{\left((\lambda_1 t)^i (\lambda_2 t)^{k-i}\right)}{i!(k-i)!} e^{-(\lambda_1+\lambda_2)t} \\ &= \left(\frac{\lambda_2{}^k}{k!} + \frac{\lambda_1 \lambda_2{}^{k-1}}{(k-1)!} + \frac{\lambda_1{}^2 \lambda_2{}^{k-2}}{2!(k-2)!} + \cdots + \frac{\lambda_1{}^{k-1}\lambda_2}{(k-1)!} + \frac{\lambda_1{}^k}{k!}\right) t^k e^{-(\lambda_1+\lambda_2)t} \\ &= \frac{\left((\lambda_1+\lambda_2)t\right)^k}{k!} e^{-(\lambda_1+\lambda_2)t} \end{aligned} \tag{6.13}$$

Hence, the probability distribution of process Y has a Poisson distribution. And as a result, the process is a homogeneous Poisson process with parameter $\lambda_1 + \lambda_2$.

QED

Proof of theorem 2.3.5

Independence of the arrival times follows from the property that increments over non-overlapping time intervals are independent. Consider for example arrival times $t_1,\ldots,t_n$, then

$$\begin{aligned} &P\left(t_1 \in [0,s_1),\cdots,t_n \in [0,s_n)\right) = P\left(N(s_1)=1,\cdots,N(s_n)-N(s_{n-1})=1\right) \\ &= P\left(N(s_1)=1\right)P\left(N(s_2)-N(s_1)=1 \middle| N(s_1)=1\right)\cdots \\ &= P\left(N(s_1)=1\right)P\left(N(s_2)-N(s_1)=1\right)\cdots P\left(N(s_n)-N(s_{n-1})=1\right). \end{aligned} \tag{6.14}$$

Now for the distribution of the arrival times, partition the interval *[0,T)* in *M* equidistant intervals of length Δt.

First it is shown that in the case of a single arrival time the variable is uniformly distributed over the *[0,T)* interval. Define p_{1i} as the probability that the first events ends up in the first bin. Then by definition of the Poisson process $p_{1i} = (\lambda \Delta t + o(\Delta t)) / C$, for some constant *C*. This constant is a result of the extra condition imposed by the fact that

$$\sum_{i=1}^{M} p_{1i} = 1. \tag{6.15}$$

Substitution of p_{1i} in Equation (6.15) yields

$$\begin{aligned} &\sum_{i=1}^{M} \frac{\Delta t\lambda + o(\Delta t)}{C} = 1 \\ &\Rightarrow C = M\Delta t\lambda + Mo(\Delta t) = T\left(\lambda + \frac{o(\Delta t)}{\Delta t}\right). \end{aligned} \tag{6.16}$$

Then the conditional probability density function belonging to this event is defined by

$$f_{T_1}(t) = \lim_{\Delta t\to 0} P\left(t \le t_1 < t+\Delta t\right) = \lim_{\Delta t\to 0} \frac{p_{1i}}{\Delta t} = \lim_{\Delta t\to 0} \frac{\lambda + \frac{o(\Delta t)}{\Delta t}}{T\left(\lambda + \frac{o(\Delta t)}{\Delta t}\right)} = \frac{1}{T} \tag{6.17}$$

for $t = i\Delta t$, $i=1,2,\ldots,M$. So that T_1 is uniformly distributed over *[0,T)*.

Now it is show that the second arrival time is also uniformly distributed. By independence of the arrival times and property 4 of the definition of the Poisson process, which said that the probability of two arrivals at the same time is of order $o(\Delta t)$, T_2 is also uniformly distributed over *[0,T)* except for the bin where t_1 arrived. But in the limit the bin size is of infinitesimal

size so does not effect the conditional distribution function of T_2 so that T_2 is also uniformly distributed.
Note that this argument holds for all succeeding arrival times, so that the proof is completed.

QED

Proof of corollary 2.3.6

Consider two independent HPP's with $N_1(t)+N_2(t) = m$ *and intensities* λ_1 and λ_2, then for $n = 0,1,\ldots,m$

$$P\left(N_1(t)=n \middle| N_1(t)+N_2(t)=m\right)=\frac{P\left(N_1(t)=n, N_1(t)+N_2(t)=m\right)}{P\left(N_1(t)+N_2(t)=m\right)}$$

$$=\frac{P\left(N_1(t)=n, N_2(t)=m-n\right)}{P\left(N_1(t)+N_2(t)=m\right)}$$

$$=\frac{P\left(N_1(t)=n\right)P\left(N_2(t)=m-n\right)}{P\left(N_1(t)+N_2(t)=m\right)}$$

$$=\frac{\frac{(\lambda_1 t)^n}{n!}e^{-\lambda_1 t}\frac{(\lambda_2 t)^{m-n}}{(m-n)!}e^{-\lambda_2 t}}{\frac{\left((\lambda_1+\lambda_2)t\right)^m}{m!}e^{-(\lambda_1+\lambda_2)t}}$$

$$=\binom{m}{n}\left(\frac{\lambda_1 t}{\lambda_1 t+\lambda_2 t}\right)^n\left(\frac{\lambda_2 t}{\lambda_1 t+\lambda_2 t}\right)^{m-n}$$

Which follows from the definition of the conditional distribution and independence of the processes.

QED

Proof of theorem 2.4.2

The proof is identical to the proof of theorem 2.3.2 and hinges on equality of the moment generating functions of the inhomogeneous Poisson process and that of the Poisson distribution as laid out in Equation (2.6).

Recall that the moment generating function of the Poisson distribution with parameter μ is given by

$$\sum_{i=0}^{\infty} e^{zi} \frac{(\mu)^i}{i!} e^{-\lambda t} = e^{\mu\left(e^z - 1\right)} \tag{6.18}$$

And for the Poisson process the proof is similar and only the last two lines of Equation (6.8) need to be replaced. Starting from third line the moment generating function for the inhomogeneous Poisson process becomes

$$\begin{aligned} &\frac{dM_{N(t)}(z)}{dt} = M_{N(t)}(z)\lambda\left(e^z - 1\right) \\ &\Rightarrow \log\left(M_{N(t)}(z)\right) - \underbrace{\log\left(M_{N(0)}(z)\right)}_{=0} = \int_0^t \lambda(\tau)\, d\tau \left(e^z - 1\right) \\ &\Rightarrow M_{N(t)}(z) = e^{\int_0^t \lambda(\tau) d\tau \left(e^z - 1\right)} \end{aligned} \tag{6.19}$$

Thus by choosing the appropriate value for μ, i.e.

$$\mu = \int_0^t \lambda(\tau)\, d\tau \tag{6.20}$$

the moment generating functions of the distributions are the same and hence the distributions themselves are equal.

QED

Proof of theorem 2.4.3

Let *{N(t): t ≥ 0}* be a NHPP with parameter *λ(t) > 0* and *t > s ≥ 0*. Consider the following moment generating function

$$M_{N(t)-N(s)}(z) = E\left[e^{z\left(N(t)-N(s)\right)}\right] = \frac{E\left[e^{zN(t)}\right]}{E\left[e^{zN(s)}\right]} \tag{6.21}$$

Then by applying theorem 2.4.2a to Equation (6.21) this results in

$$M_{N(t)-N(s)}(z) = \frac{e^{\int_0^t \lambda(\tau) d\tau \left(e^z - 1\right)}}{e^{\int_0^s \lambda(\tau) d\tau \left(e^z - 1\right)}} = e^{\int_s^t \lambda(\tau) d\tau \left(e^z - 1\right)} \tag{6.22}$$

QED

Proof of theorem 2.4.4

Consider two independent NHPP's, X_1 and X_2, with intensities $\lambda_1(t)$ and $\lambda_2(t)$, respectively. By writing out the probability distribution for the superimposed, Y, it is shown that Y is also a NHPP with intensity $\lambda_1(t) + \lambda_2(t)$.
Writing out the probability that k events occur in time interval $[0,t)$ leads to

$$P(Y=k)=\sum_{i=0}^{k}P(X_1=i,X_2=k-i)=\sum_{i=0}^{k}P(X_1=i)P(X_2=k-i). \tag{6.23}$$

by conditioning on X_1 and independence. Define the cumulative intensities over the interval $[0,t)$ as $M_1(t)$ and $M_2(t)$, that is to say

$$\begin{aligned} M_1(t)&=\int_0^t \lambda_1(\tau)d\tau, \\ M_2(t)&=\int_0^t \lambda_2(\tau)d\tau. \end{aligned} \tag{6.24}$$

Then substitution of Equation (6.24) in Equation (6.23) yields

$$\begin{aligned} &\sum_{i=0}^{k}\frac{(M_1(t))^i}{i!}e^{-M_1(t)}\frac{(M_2(t))^{k-i}}{(k-i)!}e^{-M_2(t)}=\sum_{i=0}^{k}\frac{\left((M_1(t))^i(M_2(t))^{k-i}\right)}{i!(k-i)!}e^{-(M_1(t)+M_2(t))} \\ &=\left(\frac{M_2(t)^k}{k!}+\frac{M_1(t)M_2(t)^{k-1}}{(k-1)!}+\frac{M_1(t)^2M_2(t)^{k-2}}{2!(k-2)!}+\cdots+\frac{M_1(t)^{k-1}M_2(t)}{(k-1)!}+\frac{M_1(t)^k}{k!}\right)e^{-(M_1(t)+M_2(t))} \\ &=\frac{\left((M_1(t)+M_2(t))t\right)^k}{k!}e^{-(M_1(t)+M_2(t))} \end{aligned} \tag{6.25}$$

Hence the probability distribution for process Y is a Poisson distribution with parameters $M_1(t) + M_2(t)$.

QED

Proof of theorem 2.4.5

The proof is a slight adjustment to the proof of theorem 2.3.5. Independence of the arrival times follows from equation (6.14).

Again the $[0,T)$ interval is divided into the M bins of equal size. The probability of 1 event occurring into bin i is in this case is approximated by $p_{1i} = (\lambda((i-\tfrac{1}{2})\Delta t)\Delta t + o(\Delta t))/C$. In this case C can be approximated by the midpoint rule and the assumption that the probabilities should sum to one, i.e.

$$\sum_{i=1}^{M} \frac{\lambda\left(\left(i-1/2\right)\Delta t\right)\Delta t + o\left(\Delta t\right)}{C} = 1$$
$$\Rightarrow C \approx \sum_{i=1}^{M} \lambda\left(\left(i-1/2\right)\Delta t\right)\Delta t + Mo\left(\Delta t\right). \tag{6.26}$$

Thus the conditional distribution becomes

$$f_{T_1}\left(t\right) = \lim_{\Delta t \to 0} P\left(t \le t_1 < t + \Delta t\right) = \lim_{\Delta t \to 0} \frac{p_{1i}}{\Delta t} = \lim_{\Delta t \to 0} \frac{\lambda\left(t\right) + \frac{o\left(\Delta t\right)}{\Delta t}}{\sum_{i=1}^{M} \lambda\left(t\right) + M\frac{o\left(\Delta t\right)}{\Delta t}} = \frac{\lambda\left(t\right)}{\int_0^T \lambda_1\left(\tau\right) d\tau}. \tag{6.27}$$

QED

Proof of proposition 5.2.7

Proof of part a: In order for the log partial likelihood function to be convex its second derivative should be negative. In this part of the proof it is shown that the second derivative of the log partial likelihood is strictly negative, for every value. Let B be fixed, moreover take γ unequal of zero. Then the second derivative in the direction of γ is given by

$$l'' = \frac{d^2 L}{dv^2}\left(B + v\gamma\right)\Big|_{v=0} < 0$$

Substituting these terms in equation (5.20) leads to

$$l'' = -\int \left[\left\{ E_Z\left[Y^2 U\right] E_Z\left[U\right] - E_Z\left[YU\right] E_Z\left[YU\right] \right\} E_Z\left[U\right]^{-2} E_Z\left[MGS\right] \right] dt$$

where

$$Y = \gamma^T Z \tag{6.28}$$

and E and U given by equation (5.17) and (5.19) respectively.

Note that the last expectation is proportional to the life expectation of an individual given that he is not censored and is thus positive; in addition U is also positive. So, by applying Cauchy-Schwarz's theorem for square-integrable functions in Euclidian space on the term between the brackets { }, it can be concluded that the term is positive.
For the term not to be zero it has to be shown that Y is not constant on the range where G, equation (5.25), and S, equation (5.2), are not zero after the time transformation, which is the case by assumption (H2). Furthermore by the assumption the assumption of proper parameterization (H1), it is known that γ could not be orthogonal to Z.

Thus the second derivative is strictly negative and hence the log partial likelihood function is convex.

Proof of part b: Adapt the following notation

$$I_{ij} = -D_{ij}^2 L = \int Q_{ij}(t) R(t) dt \tag{6.29}$$

Where R is defined by

$$R = E_Z [U]^{-2} MGS \tag{6.30}$$

Let Z'' denote an independent copy of Z, if f is a function taken at Z, then f'' denotes the same function taken at Z''. Note that $EE''[fg'']$ is to be interpreted as $E[f(Z)g(Z'')]$ and by independence is equal to the product $E[f(Z)]E[g(z)]$.
Using this notation write Q_{ij} as

$$2Q_{ij} = E_Z E_{Z''} \left[(Z_i'' - Z_i)(Z_j'' - Z_j) UU'' \right] \tag{6.31}$$

Substituting the terms for U results in

$$2Q_{ij} = E_Z E_{Z''}'' \left[(Z_i'' - Z_i)(Z_j'' - Z_j) GG'' \exp\left(B^T (Z + Z'')\right) SS'' \right] \tag{6.32}$$

Since all components of Z and Z'' are independent, the expectation does not alter if components from Z and Z'' are interchanged. Therefore let s be the vector of $(s_1,\ldots,s_k)$. Where the s_i's are – 1 if Z_i and Z_i'' are interchanged and + 1 otherwise. Let V be the vector $(Z''- Z)$. Note that the original vector $(Z''- Z)$ is equal to sV, where the multiplication is done component wise, i.e. $s_i V_i$.
The transformation GG'' above does not alter Equation (6.32) because of the sharpened version of the product property of the survivor function of the censoring (H'4). $(Z_i''-Z_i)(Z_j''-Z_j)$ becomes $s_i s_j V_i V_j$ and rewrite as follows.
SS'' as $\exp\{-w\cosh(b^t SV)\}$ where $w=2t\exp\{b^T(Z+Z'')/2\}$. Note that w is positive because time and exponent are positive by definition.

Let E_s denote the expectation with respect to the law on s. Note, this is a conditional expectation with respect to Z and Z''. Equation (6.32) then becomes

$$2Q_{ij} = E_Z E_{Z''}'' \left[GG'' \exp\left(B^T (Z + Z'')\right) "q_{ij} \right] \tag{6.33}$$

where

$$q_{ij} = E_s \left[s_i s_j V_i V_j \exp\left(-w\cosh\left\{b^T sV/2\right\}\right) \right] \tag{6.34}$$

Because all covariates could be interchanged it is sufficient to show the case of $i = 1$ and $j = 2$. In case of b1 for $k = 2$, q_{12} is given by

$$q_{12} \propto V_1 V_2 \left[\!\left[\exp\left\{-w\cosh\left[\left(b_1 V_1 + b_2 V_2\right)/2\right]\right\} - \exp\left\{-w\cosh\left[\left(b_1 V_1 - b_2 V_2\right)/2\right]\right\} \right]\!\right] \quad (6.35)$$

because the cosh function is even, that is to say cosh*(x)* = cosh*(-x)*. Note that q_{12} is always negative: for instance, if $V_1V_2>0$ then $|b_1V_1+ b_2V_2|>|b_1V_1- b_2V_2|$, since b is strictly positive. And since the cosh function is increasing on R^+ the term between the big brackets is negative and so q_{12} is also. If on the other hand $V_1V_2<0$ then $|b_1V_1+ b_2V_2|>|b_1V_1- b_2V_2|$ and the term between the brackets as a result is positive, making q_{12} negative again. So Q_{12} is negative, thus $-D_{12}L<0$ as was required.

In case of b2 so that $k > 2$, adopt the notation $C_1 = b_1V_1$, $C_2 = b_2V_2$ and $C_3 = b_3V_3 + \ldots + b_kV_k$. Furthermore, let s'_3 be another sign independent of s. Rewrite Equation (6.35) as

$$q_{12} = E_{s,s'_3}\left[s_1 s_2 s'_3 V_1 V_2 \exp\left\{-w\cosh\left[\left(s_1C_1 + s_2C_2 + s'_3C_3\right)/2\right]\right\}\right] \quad (6.36)$$

So, by setting

$$f\left(x\right) = \exp\left\{-w\cosh\left(x\right)\right\} \quad (6.37)$$

and integrating over s_1,s_2 and s'_3 results into

$$4q_{12} = E_{s_3\cdots s_k}\left[V_1V_2\left\{\left[\!\left[f\left(x_1\right) + f\left(x_2\right)\right]\!\right] - \left[\!\left[f\left(x_3\right) + f\left(x_4\right)\right]\!\right]\right\}\right] \quad (6.38)$$

where $2x_1 = C_1 + C_2 + C_3$, $2x_2 = C_1 + C_2 - C_3$, $2x_3 = C_1 - C_2 + C_3$ and $2x_4 = C_1 - C_2 - C_3$. Note that q_{12} is negative: for $V_1V_2>0$ the following holds $|C_1 + C_2|>|C_1 - C_2|$. And since $f(x + C_3/2) + f(x - C_3/2)$ is as a function of x even and strictly concave, so that the first term between the big brackets is bigger than the second term between the big brackets making q_{12} negative. If on the other hand $V_1V_2<0$ the following holds $|C_1 + C_2|<|C_1 - C_2|$. So that along the same reasoning the term between the { } is positive, making again q_{12} negative.
So, if the above holds than Q_{12} is also negative completing the proof.

But the function f also depends on t because w does. And it is noted that f is concave only on a neighborhood of $t = 0$. From calculus it is known that a function is concave only if its second derivative is negative. Working out the second derivative, the following should hold in order for f to be concave

$$w\left[\sinh\left(x\right)\right]^2 \le \cosh\left(x\right) \quad (6.39)$$

So, reminding that $w= 2tEE''$ and that $x = C_1 \pm C_2 \pm C_3$ the equality changes into

$$2tBB'' \le \frac{\cosh\left[\left(C_1 \pm C_2 \pm C_3\right)/2\right]}{\sinh\left[\left(C_1 \pm C_2 \pm C_3\right)/2\right]} \quad (6.40)$$

for all possible choices of signs ±. Put $D = \exp(b^TsZ)$ and $D'' = \exp(b^TsZ'')$. Notice that E and E'' have also changed to D and D'' respectively. So that by rewriting equation (6.40) above leads to

$$t \leq \left(D + D''\right)\left(D - D''\right)^{-2} \tag{6.41}$$

Thus the lower bound for the right-hand-side is the formula for the t_0 proposed. So, if $P(C<t_0)=1$ them Q_{12} is strictly negative and the result holds.

Proof of part c: Let all components be different from zero. If both i and j are taken negative the proof still holds because of the symmetry of the cosh function around *0*. Moreover when either i or j is negative the changes in sign work out nice and the result is such that q_{12} is positive. So, the first part of the proof holds.
For the second part, note that part a of the proposition applies making sure D^2_{ii} is positive. If on the other hand b_i or b_j is zero, because of symmetry of the cosh function around zero the term q_{12} is also zero. And since the place of the covariates in the vector is chosen randomly it holds for all b_i and b_j.

QED

Literature

Abrahamowitz, M. “Bias due to aggregation of individual covariates in the Cox regression model”. American Journal of Epedimiology, Vol 160, No 7, pp 696-706, 2004.

Ansell, JI and Philips, MJ. “Practical problems in the statistical analysis of reliability data”. Applied statistics, Vol 38, No 2, pp 205-247, 1989.

Ansell, JI and Philips, MJ. “Practical aspects of modeling of repairable systems data using proportional hazards models”. Reliability Engineering and System Safety, Vol 58, pp165-171, 1997.

Bretagnolle ,J and Huber, C. “Effects of omitting covariates in Cox’s model for survival data”, appearing in Scand J Statist 15: 125-138, 1988.

Cooke, RM and Morales-Napoles, O. “Competing risk and the cox proportional hazard model”. J. Stat. Plan. & Inference, 2004.

Cox, DR. “Regression models and life tables”. Royal Statistical Society, series B, Vol 34, No 2, pp 187-220, 1972.

Cox, DR and Snell, EJ. “A general definition of residuals”. 1968.

Gerds, TA. “On functional misspecification of covariates in the Cox regression model”. Biometrika, Vol 88, No2, pp572-580, 2001.

Goel, S and Bush, SF. “Biological models for security of virus propagation.” Unpublished article, 2005.

Kalbfleish, JD and Prentice, RL. “The statistical analysis of failure time data”. Wiley New York, 1980.

Kalbfleish, JD and Prentice, RL. “The statistical analysis of failure time data SE”. Wiley New York, 2002.

Lawless, JF. “Regression methods for Poisson process data”. Applied Statistics, Vol 82,No 399, pp 808-815, Sep 1987.

McCumber, J. “Assessing and managing security risk in IT systems: a structured methodology”. Auerbach publications, 2005.

Miles, J and Shevlin, M. “Applying regression and correlation”. Sage publications, 2001.

Nichols, RK and Ryan, DJ and Ryan, JJCH. Defending your digital assets: against hackers, crackers, spies & thieves. McGraw-Hill, 2000.

Nenadic, Z. http://robotics.caltech.edu/~zoran/Research/poisson/node1.html, 2005.

Rathbun, SL. “Estimation of poisson intensity using partially observed concomittant variables”. Biometrics, Vol 52, No 1, pp 226-242, March 1996.

Ryan, JJCH, Woloschek, J and Leven, B. Defense Intelligence Journal, complexities in conducting information warfare. Vol 5. Nr 1. Spring 1996.

Ryan, JJCH and Ryan, DJ. “Expected benefits of information security investments”. Unpublished paper. The George Washington University, 2005:1.

Ryan, JJCH and Ryan, DJ. “Proportional hazards in information security”. Unpublished paper. The George Washington University, 2005:2.

Ryan, JJCH. “Is Microsoft a threat to national security? The effect of technology monocultures on critical infrastructure.” Unpublished paper, 2005.

Schoenfeld [1982]

Snort Development Team. http://www.snort.org, 2005

Thommes, RW and Coates, MJ. “Modelling virus propagation in a Peer-to-peer network”. Unpublished paper, 2005

Therneau, MT and Grambsch, PM. “Martingale-based residuals for survival models”. Biometrika, Vol 77, 1, pp147-160, 1990.

Usmani, RA.“Applied Linear Algebra”, 1987.

Wang, Y and Wang C. “Modeling the effects of timing parameters on virus propagation”. Unpublished paper, 2005.

Wehner, S. “Analyzing worms and network traffic using compression”. Centrum voor Wiskunde en Informatica. April 12, 2005

Whitman, ME and Mattord, HJ. “Principles of information security”. Thomson course technology, 2003.

Wikipedia. http://en.wikipedia.org/wiki/Main_Page, 2005.