Full Research Plan

Andreas Löf

January 15, 2009

Contents

Chapter 1

Introduction

There are many problems with owning and maintaining a large scale network. Beyond a certain point they will be very costly and very difficult to operate the network. A solution to this is to make the network autonomous and allow the network do detect faults in itself.

The first section discuss the problems with large scale networks and the challenges posed by making them autonomous. The second and final section contains an outline of the rest of this document.

1.1 Problem Description

Large scale computer networks are very complex systems. As the size of a network increases, the resource demand to manage the network and cope with changes grows significantly[22].

In order to make it easier to deal with these large networks the owners usually deploy a number of monitors. The purpose of deploying these monitors is so that the operators can supervise the network and discover problems. These monitors tend to aggregate data and do not perform any kind of fault analysis or detection, leaving it to the human operators to identify the problem in the network.

There are also problems that some of these monitors do not discover. An example of one such problem is if there is a small amount of corruption of packets on a link. TCP will be able to compensate for this error without the user realising something is wrong. However, increased retransmission of corrupted packets by TCP to compensate will result in reduced link performance.

The owners of the network will eventually reach the point where they must employ a large staff to maintain and to monitor the network. The staff might also be slow to detect or deal with a problem that occurs in the network. One approach to both relieving the staff and having a faster response is to create an autonomous network.

An autonomous network is a network that is self-configuring, self-healing,

self-optimising and self-protecting. This means that the network is to some extent aware of itself in order to fulfil these goals.

One step towards this awareness is if the network can accurately identify when a problem occurs in it. In order to be able to identify potential problems within a network, we have to be able to detect events within the network. These events can be collected through both passive and active event detection methods.

Passive methods listen to network traffic without adding to it themselves whereas active methods measure the network metrics by sending probes into the network.

Listening to network traffic is very expensive in terms of resources if we do deep packet inspection. A different and more feasible approach is to inspect the flows of data instead. This does however place additional requirements on the event detection methods.

If a network is able to identify events and deduce the underlying causes of events it can then take appropriate actions to protect and heal itself.

The challenge lies in making the network infer the actual cause of the problem. The traditional approach is to use heuristics and hard coded values that trigger alarms in the monitoring systems.

I am investigating if it is possible to gain a higher detection accuracy and more information about an event when artificial intelligence is used to analyse the patterns of events generated by the event detection methods instead of using the traditional approach.

1.2 Document Outline

Chapter 2 describes the the key concepts in more detail and the related work of my research. Chapter 3 contains my hypothesis, my two main research questions as well as the approach that I will take to answer these questions. Chapter 4 contains my thesis outline and Chapter 5 contains a general time plan that I intend to follow. The final chapter, Chapter 6 contains the description of the resources that I will need and my ethics statement. Appendix A contains the FRST documents describing how data will be treated and the approval letter from the Ethics Committee.

Chapter 2

Concepts and Related Work

The first part of this chapter describes what artificial intelligence is, then network flows and network events. Active and passive network measurement techniques are then described.

The second half gives an overview of the related work, focusing mainly on event detection methods and data fusion. It shows that the research that will be undertaken is original work while closely relating to other research fields.

2.1 Basic Concepts

The concepts explained in this section are important to be familiar with in order to understand the nuances of my research. It is important that the reader notes that I am using a narrower definition of artificial intelligence than the general broad definition.

2.1.1 Artificial Intelligence

Artificial intelligence is the field of computer science that attempts to make a computer show intelligence. Intelligence as I will use it in my research is the ability to react rationally and to act rationally as defined by Russel and Norvig[19].

Rationality in the context of my research is not the ability to be Turing complete. It is rather the ability to learn from ones surroundings and to infer additional knowledge based on what happens in the surroundings.

The field contains a number of subfields, of which machine learning and inference are of particular interest.

Machine learning is the subset of artificial intelligence research where a computer is taught by example. It allows the machine to learn and adapt to new situations. For example when a machine learns to classify the weather into good and bad based on examples containing information about temperature, amounts of clouds and wind speed.

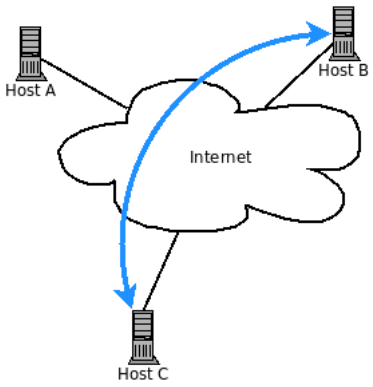

Figure 2.1: An example network showing a flow of data.

Inference is the ability of a system to reason around facts and to draw conclusions based on these facts. It is not to be confused with the ability of machine learning to learn conclusions based on examples. An example of inference is the following statements: A is a subset of B. C is a member of A. The machine can then draw the conclusion that C is a member of B.

2.1.2 Network Flow

A network flow is a flow of data between two hosts in a network for a single application. The flow is distinguished by its endpoint ports and IP addresses. An IP address is a unique identifier of computer that is publicly connected to the Internet. Each computer has 65535 ports that are used to distinguish between the applications exchanging data over the Internet. Figure 2.1 shows a flow of data between hosts B and C. This flow could be Host B downloading a file from a webserver that is running on Host C. A webserver is generally running on port 80 and a web browser gets a port allocated by the operating system when it initiates a connection to the webserver.

2.1.3 Network Event

A network event is when something unusual happens in a network. It does not have to be detrimental in its nature, but it could be. An example of a non-detrimental event would be a website gaining in popularity and quickly attracting a lot of traffic. An example of a detrimental event is when one of the links in the network fails, possibly as the result of the hardware failure or a denial-of-service attack.

2.1.4 Network Measurements

Network measurements are made in order to measure the performance and status of the network, to find problems inside the network and to collect statistics about the network usage. There are two different kinds of measurement technique, active and passive.

Active measurements have an impact on the network that they are performed on. They add to the traffic inside the network and cannot collect any information of what kind of traffic occur inside the network. Instead the active measurements are used to verify that a link, router or service is working as expected. Active methods are financially cheap to deploy and can easily be deployed across an entire network. Examples of active measurement techniques are ping, traceroute.

Passive measurement methods are used to inspect all of the network traffic flowing through each measurement point. The passive methods do not add to the network traffic or have an impact on the network at all. Passive measurement techniques are costly to deploy in a network, both in computational and financial terms and are thus not widely deployed over an entire network. Examples of passive measurement techniques are libpcap[16] and DAG hardware capture cards from Endace[7].

There are two types of passive network monitoring; deep packet inspection and flow inspection. Deep packet inspection examines the payload of every packet passing the monitor and is thus extremely resource intensive. Flow inspection only looks at the network flows, it does not suffer from the same resource requirements as deep packet inspection but it also produces less information.

2.2 Related Work

This section gives an overview over the work that relates to my planned research.

The state of the art of network monitoring is generally a combination of tools like Cacti[3], Nagios [18], Ping, Traceroute, iperf[11] and tcptrace[23]. These tools are generally implemented using a combination of active and passive measurements to form a monitoring environment of the network.

These tools do not attempt to find the cause of any problems in the network and tend to only present aggregated information to the operators. Autonomous networks deal with the information on a more detailed level and can identify the problems themselves.

Autonomous networks are not a new concept. There have been a number of papers written about the concept [5, 8, 25, 21, 22]. All of these papers either implicitly or explicitly discuss different artificial intelligence techniques.

My research focuses on determining the methods to detect network events, combining the events to gain more information and investigating what types of artificial intelligence are suitable. My research does not focus on the self repairing aspects or decision making aspects of autonomous networks.

2.2.1 Event Detection Methods

As I stated in my problem description, autonomous networks need to be able to detect events. I have primarily investigated event detection methods in network flows. The fact that a malicious network flow can be distinguished from a non-malicious flow [1] paved the way for a number of event detection methods [13, 14] that purely use network flows.

These methods all detect anomalies in the flow of network data, which is useful for detecting both attacks on the network as well as changes in traffic patterns due to faults in the network. They do however not provide the possibility of detecting exactly where in the network the fault lies or what the cause might be unless they are deployed over every link in the network, which is unfeasible due to the associated costs of traffic monitoring.

In order to more efficiently pinpoint the faults in the network,active monitoring techniques can be used. One of the more interesting papers using active measurement was published by Logg et al.[17]. They present a method that analyses the available bandwidth over links to detect events by finding significant unexpected changes. The idea is similar to the the one that Brutlag suggested[2], but the difference is that Brutlag uses Holt-Winters whereas Logg et al. described a completely new method.

Cottrell et al.[4] have published a paper that evaluates these two methods as well as the Kolmogorov-Smirnov test, Mark Burgess Technique and a combination of some of these techniques.

All of the techniques that have been discussed so far have one flaw in common, they poorly adapt to slow changes in traffic patterns. Gu et al.[10] created a new method that borrows from the field of machine learning to detect anomalies. This method looks at the entropy of the data that passes through a link, which means that it is a passive method that looks at the flow of data through the network. This method is however computationally expensive and requires a significant amount of event free training data.

I want to attempt to combine the events generated from these methods since no method can detect all kinds of network events and they will not necessarily detect the same events. Càrdenas et al.[6] have published a paper where they discuss the different methods of measuring accuracy of intrusion detection systems as well as a framework for evaluating intrusion detection systems. It is a very important paper since my research will have the same difficulties with measuring performance the intrusion detection researchers have.

2.2.2 Artificial Intelligence Approaches to Autonomous Networks

Koks and Challa [12] have written a technical report that introduces two of the main techniques used for data fusion, Baysian and Dempster-Shafer. They also prove an overview of the applications data fusion, which shows that data fusion is suited for my problem.

Data fusion has been introduced into intrusion detection systems and there

are two papers that are of particular interest for the research in this plan. Widder et al. [26] introduced the concept of detecting unknown patterns of events amongst a more general set of events. Gu et al.[9] published a paper in which they compare ensembles of intrusion detection systems and describe a method to fuse the events generated by the intrusion detection systems. This is similar to what I wish to achieve, but my focus is on detecting faults in the network instead of intrusions.

Another approach to autonomous networks is to use expert systems. Vlahavas et al.[24] describes a system that they have developed. The system is a decentralised expert system, with each node dealing with local problems and the entire system dealing with wider network problems. The system is however limited to dealing with situations that it has encountered before and being taught how to deal with them.

2.2.3 Conclusions

The research that will be conducted in thus project focuses on network monitoring, not network intrusion detection systems. The major difference is that I will not focus on attacker data but rather on network infrastructure monitoring and surveillance. The key components are the event detection methods, the methods for combining events and the artificial intelligence techniques. To my knowledge, nobody has done a substantial amount of prior work to bring these fields together with the explicit purpose of finding network anomalies, not just detecting network intrusions.

Chapter 3

Research Questions

Chapter 1 outlined the problem with having large scale networks and managing them. The amount of events generated by the network will be too many for the human operators to deal with. Instead, artificial intelligence can be used to handle this events.

For my research, I have established a hypothesis and in order to verify the hypothesis I will investigate what types of artificial intelligence is suitable and if it is more effective to fuse data from several event detection methods. The chapter is divided into two main parts, the first discuss the research questions and the hypothesis and the second the approach that I will take.

3.1 Hypothesis

The hypothesis that I am investigating is:

Using artificial intelligence to combine the outputs of several event detection methods will result in both higher accuracy and more output compared to using the event detection methods individually.

More output means that it is possible to infer knowledge from the combination of the outputs of the event detection methods that is impossible to obtain from the event detection methods individually.

The measures that are used in the comparison will be the percentage of false positives/negatives and true positives/negatives for each method that I compare. Each method will be evaluated with the same data set in order to ensure that the results will be comparable.

3.2 Research Questions

This section describes my two main research questions. The questions must be answered in the order stated since the second question depends on the first.

3.2.1 What Types of Artificial Intelligence Are Suitable?

There are many different concepts and techniques used in the field of artificial intelligence. Not all of them are suited to deal with the large amounts of data that are generated within a network. It is also important that the techniques used can deal with unknown data and adapt over time since the usage of the network will change over time.

I will answer this question by doing a survey of the field.

It is impossible to foresee all possible events that can happen in a network, which makes it very important that the artificial intelligence techniques can deal with unknown data in such a way that it will not be misclassified.

Training and classification speed are both important, but for different reasons. Being able to train the artificial intelligence tools quickly is important so that the tools can start to process previously unknown data as fast as possible. Rapidly classifying data is a requirement if the techniques will be used in real time to look at network events.

Section 3.3.4 describes how I will undertake the survey.

3.2.2 Is It More Effective to Fuse Data From Several Event Detection Methods?

This question establishes whether the hypothesis can be upheld or not.

The two main criteria for my investigation are the accuracy and the possibility to infer knowledge from combinations of events.

I will measure the accuracy by comparing the percentage of correctly classified events and the percentage of misclassified events. A misclassified event is either a false negative or a false positive.

Section 3.3.5 describes the approach that I will use to answer this question.

3.3 Approach

There is a certain amount of work that I need to carry out before I can start to investigate the research questions. This section outlines the largest parts of that work.

3.3.1 Creating Pre-classified Training Data

I cannot compare any methods until I have pre-classified data. This means that I will have to examine a sizable data set and classify it by hand in order to have known data that I can use to evaluate my research on.

I will work together with Yu Wei Wang on this part of my research since the both of us need to create pre-classified data.

This will be done by choosing a time period of two weeks from the Waikato trace set available on the WITS archive[27]. It is important that the time period is within either of the teaching semesters at the university when the usage patterns are normal. Two weeks was chosen because we will have seasonal

variation within both every day of the week and the week itself. Having two weeks of data means that we will have one week of training data and one week of evaluation data. In addition to this we intend to look at well known events and classify the data in and around them as well.

In order to classify the trace sets we will implement different event detection methods and manually investigate the events that these methods report. We will implement two different classes of event detection methods. One that only works on the total amount of data in the network and the second that works on a flow basis.

The specific methods that we will use that work on the total amount of network data are: Holts-Winter Exponential Smoothing[17], Kalman Filters[20] and an ARIMA model. we will combine these methods with the event detection methods described in [2, 4, 17].

The methods that we chose for the flow based network analysis are Sketch Subspaces [15] and a method based on principal component analysis that is yet to be decided.

These methods have been chosen because they are relatively simple to implement and have achieved good results on other data.

If we suspect that these methods do not capture all network events we will look into more methods until we are confident that we have discovered the vast majority of all events in our data.

3.3.2 Choosing Event Detection Methods

Before I can start to focus on my research question, I must choose a set of event detection methods that I will use during my comparisons. I will investigate different methods and see which ones are the most suitable for my research.

I will investigate both active and passive event detection methods, but any passive event detection methods I investigate must work on network flows instead of using deep packet inspection due to the limitations of deep packet inspection.

I will focus on finding several methods that can detect the same types of events using different data. This will allow us to determine if it is possible to infer more knowledge from the combination of events.

Details

I will start out by doing a paper survey and categorise the methods based on the types of events they detect and the level of data they operate on (total amount of network traffic or network flows).

I will then further categorise the event detection methods based on the types of data they discover based on types of input.

The final step in choosing the event detection methods will be to implement the most likely candidates and see how they perform in reality. If I have any doubts about my implementation I will attempt to contact the original au-

thor of the method, request their evaluation data and attempt to recreate their published results.

3.3.3 Creating a Framework For Comparing Event Detection Methods

It will be impossible to conduct my research unless I have a framework that I can use to create and compare my results.

The framework must support different event detection methods, and give us the ability to directly compare the results with the results from the artificial intelligence tools.

The details and the design of the framework is something that I will finalise when it is time to start to implement the framework.

The initial requirements placed on the framework are:

- The framework must be modular - detection algorithms must be easy to swap
- The output must be start of event and end of event as well as type of event.
- If a cause of the event is detected, the cause must be apparent.
- The framework must be able to pre-process data in order to present it to the network events.
- The output from one event detection algorithm must be usable as input for another algorithm.
- The output from one or several event detection algorithms must be routable to several outputs.
- The algorithms can output debug data during runtime.

This is not an exhaustive list of the framework and I will draw on the experiences learned by Càrdenas et al.[6] and conduct a more thorough analysis at a later time.

3.3.4 Survey of Artificial Intelligence Techniques

This section describes the approach that I will use to conduct my survey of different types of artificial intelligence.

I will primarily focus on the following criteria while conducting the survey:

- Ability to handle unknown data.
- Classification speed.
- Accuracy (rates of false positives/negatives).

I will conduct the survey by looking at the available machine learning tools and reasoning tools as well as doing a literature study. I will focus on the existing implementation of tools rather than constructing new tools.

In the field of machine learning I will look mainly at multi-class classification and one-class classification.

A part of the survey will be to establish what types of inference are suitable.

If I find a large number of suitable techniques I will look at the strengths and weaknesses of each and add more criteria until I am left with a small and manageable number of techniques.

I may encounter tools that are suitable to use to answer the question posed in Section 3.2.2 but the tools are implemented in such a manner that this is difficult. If that happens I will modify the tools to work with the framework described in Section 3.3.3.

3.3.5 Fusing Data From Several Event Detection Methods

This section describes the approach to the research question stated in Section 3.2.2.

Answering this question depends on all of the other work described in Section 3.3 being done and that I have the selection the artificial intelligence tools as defined in Section 3.2.1. The framework that is described in Section 3.3.3 will be used to perform the comparisons.

The first step will be to use the pre-classified data on the event detection methods themselves to establish the relative accuracy of each method.

The second step will be to investigate the output of the event detection methods and investigate if there is any correlation between events from several methods.

This correlation will then be used for the data fusion in order to see if I will gain a higher accuracy in the third step.

Step four will be to see if I can gain richer information by using the artificial intelligence techniques that I evaluated in order to answer the question posed in Section 3.2.1.

The final step will be to evaluate my results.

Chapter 4

Thesis Outline

The outline of my Ph.D thesis is as follows:

1. Introduction
2. Background Study
3. Technical Platform
4. Evaluation of Event Detection Methods
5. Evaluation of Artificial Intelligence Techniques
6. Analysis of Event Detection Fusion
7. Conclusions

My thesis will most likely also contain a number of appendices in the end.

Chapter 5

Timeplan

I plan to finish my Ph.D in a total of three years. The sum in the table is 30 months since I have already spent 6 months on my Ph.D studies.

It is possible that several tasks will be performed in parallel. In addition to the main research I have put in the estimates I have made for writing papers and progress reports.

Task	**Time**
Progress reports	2.5 months
Write conference papers	5 months
Create test data	2 months
Create framework	3 months
Machine learning techniques	2.5 months
Implement event detection methods	3 months
Compare aggregated event detection methods	3 months
Write thesis	9 months
Sum	**30 months**

Table 5.1: Thesis timeplan

A tentative timeplan of my intended schedule can be found in Table 5.1. Each progress report has been marked as a milestone together with the final submission of the thesis. The main tasks have been allocated 80% of my available time and paper writing 20%.

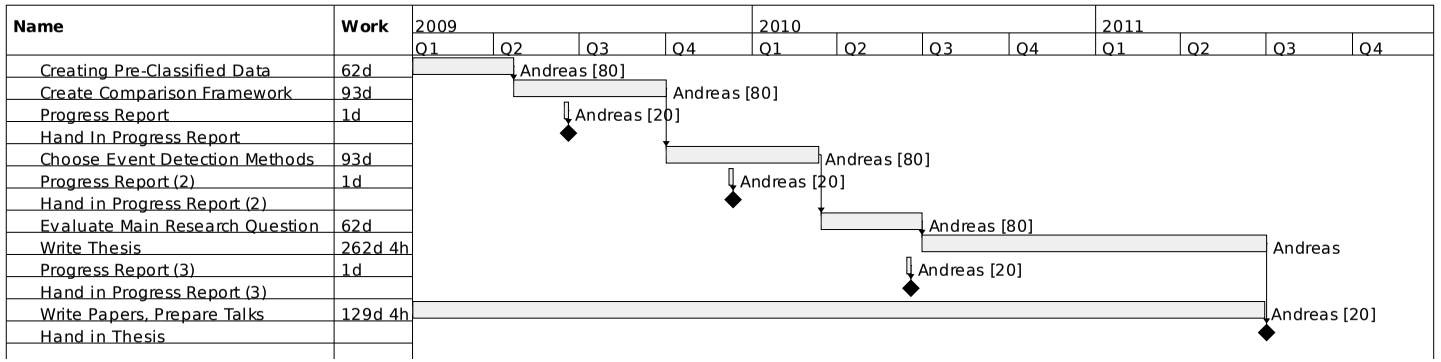

Figure 5.1: A Gantt chart displaying the main stages during this project.

Chapter 6

Other Information

This chapter lists the required resources as well as the ethical considerations of my research.

6.1 Resources

I will require access to a workplace, the library, the Internet as well as network traffic to analyse and the computational capacity to analyse the network traffic.

6.2 Ethics Statement

My primary supervisor, Richard Nelson, has ethical approval for a FRST project with the title "UOWX0705 Autonomous Networks" . My research falls within this project and is thus included in the approval. The documents describing how captured network data is treated have been attached at the end of my research plan.

I have signed the documents concerning access to the W.A.N.D repository of network traces.

Bibliography

[1] P. Barford and D. Plonka. Characteristics of network traffic flow anomalies. In *IMW '01: Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, pages 69–73, New York, NY, USA, 2001. ACM.

[2] J. D. Brutlag. Aberrant behavior detection in time series for network monitoring. In *LISA '00: Proceedings of the 14th USENIX conference on System administration*, pages 139–146, Berkeley, CA, USA, 2000. USENIX Association.

[3] Cacti. http://www.cacti.net, 12 2008.

[4] R. L. Cottrell, C. Logg, M. Chhaparia, M. Gngonev, F. Haro, F. Nazir, and M. Sandford. Evaluation of techniques to detect significant network performance problems using end-to-end active network measurements. In *Network Operations and Management Symposium, 2006. NOMS 2006. 10th IEEE/IFIP*, pages 85–94, 2006.

[5] M. Crosbie and G. Spafford. Defending a computer system using autonomous agents. In *Proceedings of the 18th National Information Systems Security Conference*, 1995.

[6] A. A. Crdenas, J. S. Baras, and K. Seamon. A framework for the evaluation of intrusion detection systems. In *Proceedings of the 2006 IEEE Symposium on Security and Privacy*, volume 0, pages 63–77, Los Alamitos, CA, USA, 2006. IEEE Computer Society.

[7] Endace. http://www.endace.com/, 2008 12.

[8] E. Gelenbe. Towards autonomic networks. In *International Symposium on Applications and the Internet (SAINT'06)*, volume 0, pages 2–7, Los Alamitos, CA, USA, 2006. IEEE Computer Society.

[9] G. Gu, A. A. Cárdenas, and W. Lee. Principled reasoning and practical applications of alert fusion in intrusion detection systems. In *ASIACCS '08: Proceedings of the 2008 ACM symposium on Information, computer and communications security*, pages 136–147, New York, NY, USA, 2008. ACM.

[10] Y. Gu, A. McCallum, and D. Towsley. Detecting anomalies in network traffic using maximum entropy estimation. In *IMC'05: Proceedings of the Internet Measurement Conference 2005 on Internet Measurement Conference*, pages 32–32, Berkeley, CA, USA, 2005. USENIX Association.

[11] Iperf. sourceforge.net/projects/iperf, 12 2008.

[12] D. Koks and S. Challa. An introduction to bayesian and dempster-shafer data fusion. Technical Report DSTOTR1436, Australian Department of Defence, 2003.

[13] A. Lakhina, M. Crovella, and C. Diot. Diagnosing network-wide traffic anomalies. In *Proceedings of the ACM SIGCOMM 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, pages 219–230, 2004.

[14] A. Lakhina, M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. *SIGCOMM Comput. Commun. Rev.*, 35(4):217–228, 2005.

[15] X. Li, F. Bian, M. Crovella, C. Diot, R. Govindan, G. Iannaccone, and A. Lakhina. Detection and identification of network anomalies using sketch subspaces. In *IMC '06: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pages 147–152, New York, NY, USA, 2006. ACM.

[16] libpcap. http://www.tcpdump.org/, 2008 12.

[17] C. Logg, L. Cottrell, and J. Navratil. Experiences in traceroute and available bandwidth change analysis. In *NetT '04: Proceedings of the ACM SIGCOMM workshop on Network troubleshooting*, pages 247–252, New York, NY, USA, 2004. ACM.

[18] Nagios. http://www.nagios.org, 12 2008.

[19] S. Russel and P. Norvig. *Artificial Intelligence A Modern Approach*. Prentice Hall, 2:nd edition, 2003.

[20] A. Soule, K. Salamatian, A. Nucci, and N. Taft. Traffic matrix tracking using kalman filters. *SIGMETRICS Perform. Eval. Rev.*, 33(3):24–31, December 2005.

[21] R. Sterritt. Towards autonomic computing: effective event management. In *Software Engineering Workshop, 2002. Proceedings. 27th Annual NASA Goddard/IEEE*, pages 40– 47, 2002.

[22] R. Sterritt and M. Hinchey. Why computer-based systems should be autonomic. In *IEEE International Conference on the Engineering of Computer-Based Systems*, volume 0, pages 406–412, Los Alamitos, CA, USA, 2005. IEEE Computer Society.

[23] tcptrace. http://irg.cs.ohiou.edu/software/tcptrace, 12 2008.

[24] L. Vlahavas, N. Bassitiades, I. Sakellariou, M. Molina, S. Ossowski, I. Futo, Z. Pasztor, J. Szeredi, I. Velbitskiyi, S. Yershov, and I. Netesin. Expernet: an intelligent multiagent system for wan management. *Intelligent Systems, IEEE [see also IEEE Intelligent Systems and Their Applications]*, 17(1):62–72, 2002.

[25] S. Wallin and V. Leijon. Rethinking network management solutions. *IT Professional*, 8(6):19–23, 2006.

[26] A. Widder, R. v. Ammon, P. Schaeffer, and C. Wolff. Identification of suspicious, unknown event patterns in an event cloud. In *DEBS '07: Proceedings of the 2007 inaugural international conference on Distributed event-based systems*, pages 164–170, New York, NY, USA, 2007. ACM.

[27] Wits. https://secure.wand.net.nz/wits/index.php, 2008 12.

Appendix A

FRST Documentation

This appendix contains the original statements concerning user privacy and how the network traces will be handled. It also contains the response from the ethics committee.