Internet Analysis System (IAS)

https://www.internet-sicherheit.de/

October 12th 2008

White Paper

AUTHORS
Malte Hesse, <hesse[at]internet-sicherheit[dot]de>
Prof. Norbert Pohlmann, <pohlmann[at]internet-sicherheit[dot]de>

Institute for Internet Security
University of Applied Sciences Gelsenkirchen
Neidenburger Str. 43
45877 Gelsenkirchen
Germany

About if(is)

The Institute for Internet Security is an independent, scientific facility of the University of Applied Sciences Gelsenkirchen in Germany. One of our overall tasks is to advance research and development in the area of Internet Security and to improve the statutory framework in this area as well. The main exploratory focus is split in four main areas, which are the (i) research on the object “Internet” and Internet Early Warning Systems, (ii) e-mail security, (iii) Trusted Computing and Trusted Network Access Control (tNAC) as well as (iv) Identity Management. Besides research and development, we aim to be a creative service provider offering prototypes and research solutions to our partners. Within this White Paper we will focus on the Internet Analysis System, a sensor technology we have developed in cooperation with the German Federal Office for Information Security. We also use this sensor technology as a basis for our work in the area of research on the Internet and Internet Early Warning Systems.

Content of the White Paper

1 Introduction

One intention of the Institute for Internet Security is to create and analyze local and above all global perspectives in order to make the generation of the global view of the Internet possible. Therefore, our team has developed a passive sensor technology, that can continuously collect statistical raw data with sensors, which are placed at selected spots of the Internet communication infrastructure. At the moment we have implemented sensors mainly in Germany to monitor the internal government network and the networks of Universities and Companies. In addition, we have also found partners in Austria and Brazil.

The raw data is captured from header information of the passing network traffic, by counting the occurrence of (currently) 870.000 different parameters. This processing is assuring that all sensitive header information - like IP addresses and user data - is left out and therefore the data is not privacy-sensitive, avoiding ethical, privacy, and legal challenges that other data-collecting systems are plagued with. Additionally, our processing is designed to have very high performance and it allows frequent transfers of the collected data to our centralized database using encryption. This enables us to collect and store data securely over a long period of time. The general approach is different from Intrusion Detection Systems, which only collect information in case of a specific exception event, and different from other monitoring systems that store highly confidential content or IP addresses, which are privacy sensitive. These mentioned systems are well proven in the local environment, but cannot cope with the global challenge.

2 Aims and Tasks of the Internet Analysis System

The task of the Internet Analysis System on the one hand is to analyze local communication data in defined sub networks of the Internet, and on the other to create a global perspective of the Internet by bringing together the large number of local views. The functions of the Internet Analysis System can be divided up into the four segments of (i) pattern formation and creation of a knowledge base, (ii) description of the actual status, (iii) alarm signaling and (iv) forecasting.

The main task of pattern formation is a comprehensive analysis and interpretation of the communication parameters of Internet traffic with the aim of detecting technology trends, interrelationships and patterns, which represent the various statuses and perspectives of the Internet. We also want to create a knowledge base, which we can use to understand the functioning of the Internet from the “communication behavior” point of view.

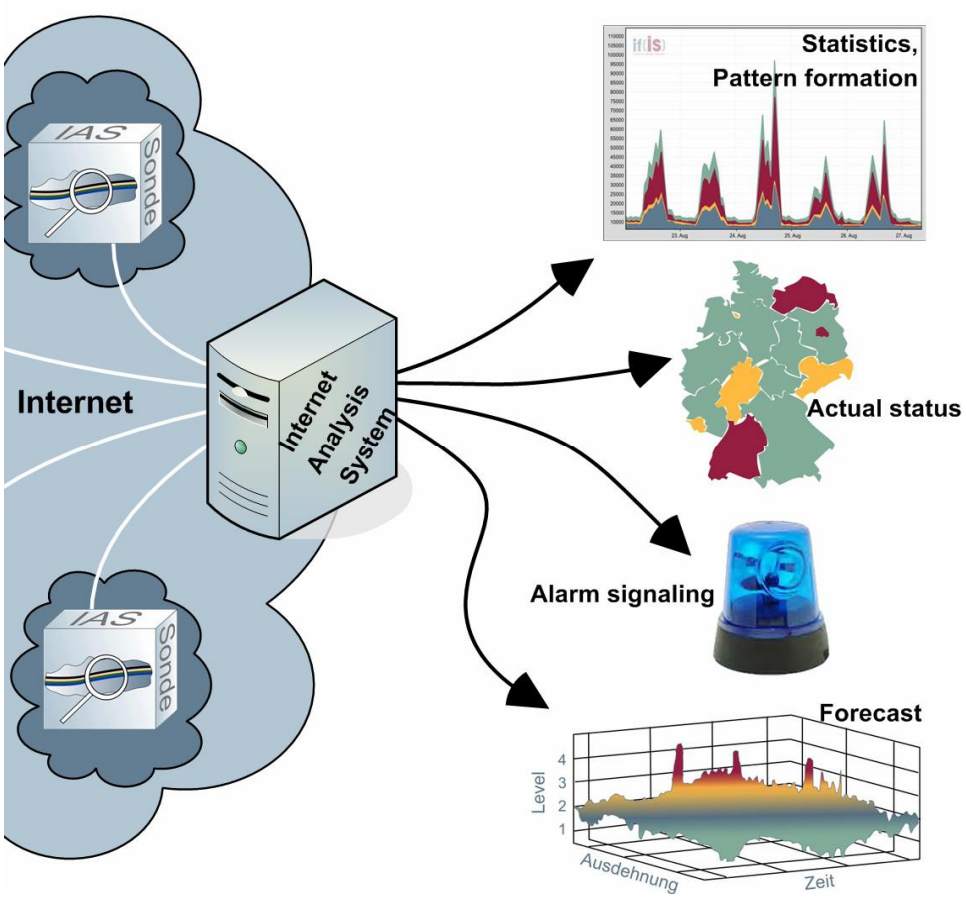

Fig. 1: Tasks of the Internet Analysis System

On the basis of this knowledge a search is carried out for anomalies among the current measured values and the causes of status changes are analyzed and interpreted. Here, it is important to find out, whether the status anomalies have a natural origin - for example as a result of a technological change - or whether they are attributable to a wanton attack.

With knowledge of the current status of a communication infrastructure and the use of historical - i.e. previously collected - information (knowledge base) it is possible in the case of significant changes to traffic volumes or communication data to generate a warning message, on the basis of which, measures can be initiated to protect and maintain the correct functioning of the Internet. By this, we can limit the damage caused by a possible attack.

A further important function is the visual depiction of the Internet status similar to a weather or traffic jam map. Here, intuitive depictions are being developed, with which the most important parameters are discernible at first glance [Spoo07] (please take a look at the example in Fig. 2).

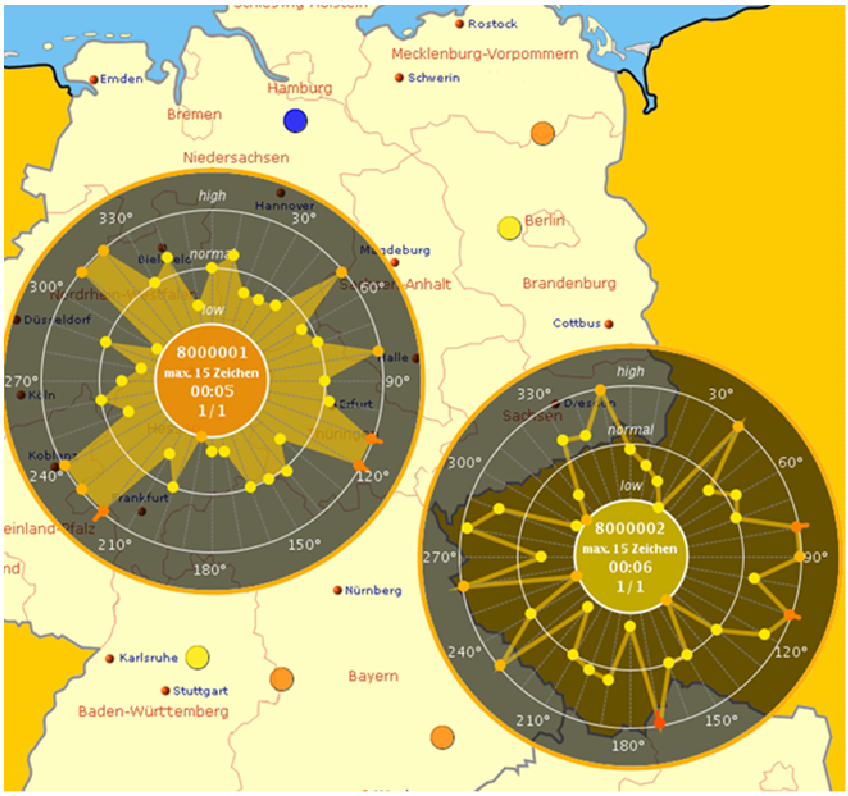

Fig. 2: Security Situation of the German Internet

Through the examination and analysis of the extrapolated profiles, technology trends, interrelationships and patterns, it will be possible, by means of an evolutionary process of the acquired results, to make forecasts of changes of the Internet status. In this manner it is possible to detect indications of attacks and important changes at an early stage and forecast the effects of the damage, which helps us to avoid this damage created by a possible attack.

3 Selection of sensor environment

The sensor can be placed in all IP-based communication infrastructures ranging from local home networks over corporate networks up to the level of autonomous systems. One question that needs to be answered is, where the sensors should best be placed for an Internet Situation Awareness. The selection of network operators as partners should of course be done due to scientific evaluations (rectangular distribution). For the future, once the technology has reached a wider level of acceptance, this will leave some unsolved complex statistical challenges, which we have to face once this opportunity arises: What is the accurate number of sensors and where exactly do these sensors need to be placed to generate a global view in a representative manner without creating to much overhead?

4 Principle of Raw Data Collection

Fig. 3 shows the principle of raw data collection by the probes. This is divided up into three sections. (i) The Internet is represented on the left. In this example packets of three different application sessions are shown: related HTTP packets, an FTP session and an SMTP

session. (ii) The probe is located in the middle of Fig. 3. The packets of the three applications are accessed passively by the probe one after the other in their random order and evaluated. The packet, that is accessed, is channeled through several analysis categories, each of which is responsible for a certain protocol. These evaluate strictly defined communication parameters in the protocol header at the various communication levels, which are not relevant to data protection laws. (iii) The counters allocated in the counting system are incremented according to how the header information of the packet is filled out. The frequency of certain header information is recorded in the same way as on a tally sheet.

Let us take a look at the following simple example: In Fig. 3 the accessing of the FTP packet is recorded by incrementing the FTP counter by 1. The raw data is therefore an aggregation of counters, i.e. counters of communication parameters that have appeared at the various communication levels over a defined period. The packet - in Fig. 3 a FTP packet - is immediately deleted physically, i.e. irreversibly and without trace, by the probe [Pohl07].

The sensor monitors for instance the protocols: IP, ICMP, TCP, UDP, HTTP, SMTP, FTP DNS, EMULE, IRC, RTP, SIP, Skype, etc [Ricc08]. For each protocol a different number of monitored parameters is implemented in the sensor technology and therefore observed by it. The number is depending on the importance and distribution of the protocol. Therefore, we count for SMTP 1.624, for HTTP 1.123, for DNS more than 9.000, for RTP only 47 and for TCP 680.551 different parameters. The number for TCP is as high, because it includes the observation of the different ports as well. These counters sum up to a total of currently 876.596 different parameters, for which the occurrence is registered by the sensor [Ricc08].

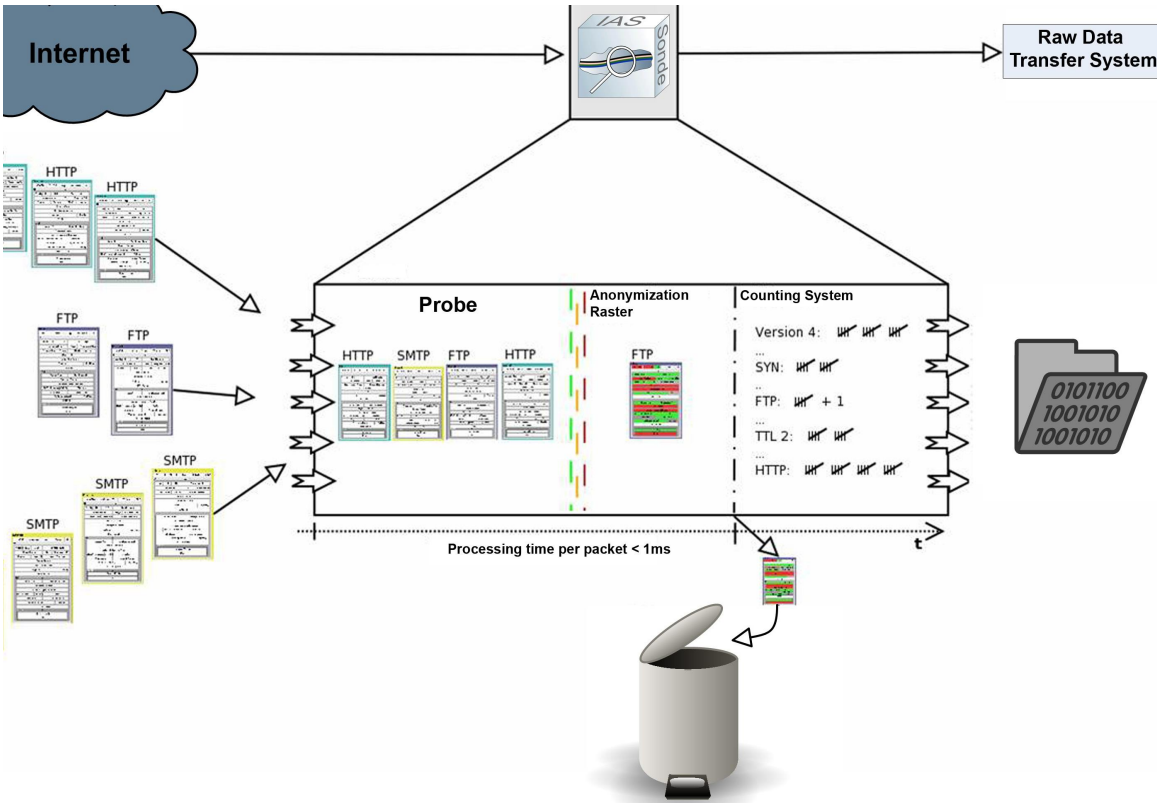

Fig. 3: Principle of raw data collection

5 Data privacy vs. data confidentiality

Speaking from our perspective, the European Union has very high standards for the privacy of the data of citizens. This was considered by the design of the sensor. Therefore, no portions of the communication packets that could be linked to an individual, is further processed or stored. This includes the payload of the packets at the level of the application layer. Even IP addresses are considered to be off limits, due to the fact that the Internet Service Provider can link the address to a customer (real person).

Reconstitution of the context of a packet or only a communication parameter is not possible or necessary. At intervals definable by the partner the counter readings (raw data) of the sensor can be transmitted securely to the raw data transfer system using encryption [Deml07]. For most partners this interval is set to be five minutes. All of this information is completely anonymous, as shown in Fig. 4.

On the right behind the colon character are the counter values for each parameter specified on the left. Each line stands for one counter. For example, line 2 indicates that 1,123,149 packets with the IP protocol number 17 (UDP) appeared in the prescribed time interval. The total table has more than 870.000 lines.

ID	*Description*		*Count*
131134	IP (Protocol Number 6)	:	18.854.151
131145	IP (Protocol Number 17)	:	1.123.149
327708	TCP (Flags: SYN)	:	334.435
327723	TCP (Flags: FIN/ACK)	:	480.697
327724	TCP (Flags: SYN/ACK)	:	275.779
545857	HTTP (Request Method POST)	:	2.026
545861	HTTP (Request Method GET)	:	293.616
545863	HTTP (Request Method HEAD)	:	18.992

Fig. 4: example of results of the counting system in the probe

One of the largest tradeoffs caused by complying with the privacy laws is the fact that we cannot work with the IP address of a potential attacker and with the payload of the packets. Therefore, we can only locate the origin of the attack roughly by locating the sensor that first recorded a phenomenon. In addition to this, we can estimate with a certain percentage, if it was a distributed or a centralized attack by using certain parameters.

And some attacks that are hidden in the user data and aim to manipulate the server like the “INVITE of Death (IoD)” for Private Branch Exchanges cannot be detected [Ricc08]. Therefore, other systems that are used by the operators of AS in case of the detection of a certain local event offer an enrichment to the Internet Analysis System and are necessary to create a continuous Internet Situation Awareness.

Of course, one could argue scenarios, in which the Internet Analysis System could also be used in a privacy violating fashion. Recently we have passed out an enhanced DSL router to students on a voluntary basis. On top of the DSL router hardware we have implemented this described sensor in combination with another development of us, an active drone monitoring the availability of services out of the perspective of users [Oste06]. Since we can link the raw

data coming from each router directly to a student, this is of course a scenario, in which we need the consent of the concerned. So in all cases, in which the sensor raw data can be linked to one single individual, we have a data privacy problem. If it is not possible to get the consent of the concerned the sensor should only be placed in working environments with at least 50 individuals. But this DSL scenario is not what the sensor was designed for. The great advantages show up, when placed in the communication infrastructure of larger networks.

So we are proud to say that we do not have to deal with privacy issues. But we are aware that by processing the raw data we might reveal information - like from economical nature - that to some extent can prove to be critical. Therefore, we have to deal with data confidentiality issues. If one links the knowledge about a crisis in a product of a company, with the fact that there is no increased level of e-mail communication on the weekends, this might show that the staff of the observed organization is not actively working on solving the problem. This is a similar situation to the DSL router scenario, but in this case it is not linked to an individual but to one single organization. We have a great understanding for the fact that the confidentiality of the raw data of individual organizations is essential for the success of the whole idea. We encounter this by establishing trust and transparency on an organizational level and by using encryption and smart processing of data on a technical level. We need to make sure that the data of an organization does not get in the hands of others.

6 Some results of the Internet Analysis Systems

For the purposes of illustration some results are presented in this section in order to provide an idea of the abilities of the current status of development of the Internet Analysis System. At present there are - as mentioned - approximately 870,000 different counters for communication parameters implemented on various layers of the communication. This large number clearly shows, how complex the results can be. Here, we now present some basic examples:

6.1 Types of e-mail messages

Fig. 5 shows the ability of the system to record the statistics of the headers of the e-mails sent via SMTP. The distribution can provide information on general communication behavior, as well as deviations from it.

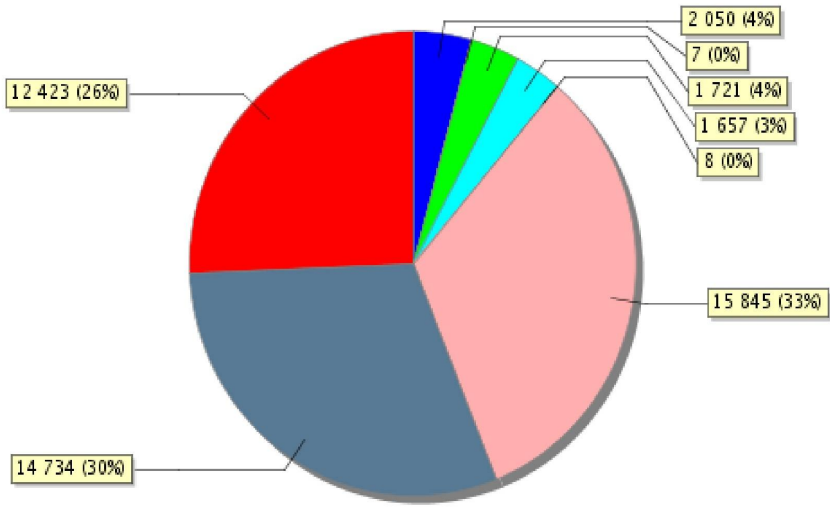

Fig. 5: Distribution of e-mail content types

Fig. 5 shows an example of normal behavior in which the total number of messages without attachments represents 60% of all messages. These e-mails include messages with the text/plain (12423), text/html (7) and multipart/alternative (14734) content types. By rule, e-mails with attachments are provided with the multipart/mixed (15845) content type. A mixed form are presented by e-mails with the multipart/related (657) content type. Here, for example, images are integrated directly into the text. If these e-mails are included in the total count of e-mails, which are having an attachment, approximately 36% of all e-mails are sent with an attachment. The remaining 4% essentially consist of confirmations of reading with the multipart/report (2050) content type. An abrupt change of these values in particular, may indicate a wave of spam affecting a company from the outside, or indicate that a computer is sending spam from within the company. It could also be an indication for an attack with malware attached to e-mails.

6.2 Transport protocol distribution

Fig. 6 shows the distribution of the protocols of the transport layer used over a period of several days for a specific communication line.

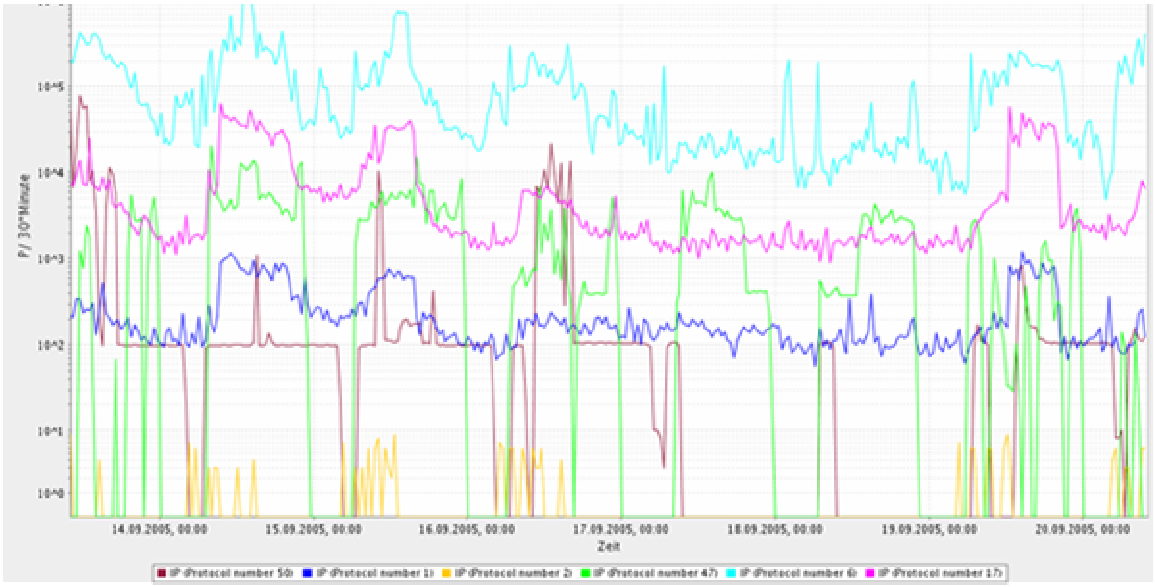

Fig. 6: Protocols of the transport layer

From past readings the Internet Analysis System has stored the profile of the standard deviation, which can be used to detect and display an indication of untypical behavior. Additionally, the use of certain protocols can be determined, enabling capacity planning for the use of Virtual Private Networks (ESP protocol), for example. Protocol dependencies can also be detected: UDP appears to be proportional to TCP, which can be attributed to the dependencies of HTTP (using TCP) and DNS (using UDP).

6.3 TLS cipher suites really used for encryption

For the secure communication between clients and servers so called cipher suites have been pre-defined consisting of methods for the key exchange with authentication and algorithms for encryption as well as for data integrity. Which cipher suite is used for the communication is then negotiated between client and server depending on availability of algorithms and set preferences. If the browser is connecting the web server the browser offers the possible crypto suites to the server. Then the web server decides, which crypto suites should be used for the communication.

Since modern encryption is based on problems of complexity and due to the fact that weaknesses in some methods have been identified, some cipher suites should no longer be used with growing performance of the available computer systems. But sometimes the use of certain very insecure cipher suites is mandatory, due to questionable national laws, which are supposed to ensure legal interception especially in so called “rogue regimes”.

The Internet Analysis System can monitor from authentic network traffic, which cipher suites are really being used. This information is very interesting for all national representatives in

charge for the monitoring of the infrastructure Internet. So far they have to base their decisions on very little available information. Most of the time they are not aware of the actual situation.

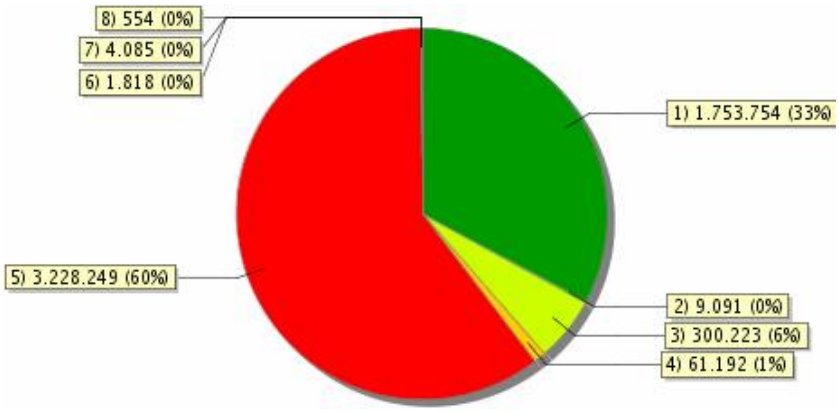

Fig. 7: Distribution of cipher suits used

In one of our monitored sub networks we have recorded the following distribution (Fig. 7): In 60% of all encrypted communication the very common RSA_WITH_RC4_128_MD5 (5) cipher suite was used, in 33% the improved and more secure DHE_RSA_AES_256_CBC_SHA (1) cipher suite and in about 6% the RSA_WITH_AES_128_CBC_SHA (3) cipher suite was used, which is from the perspective of security also fine. But we have detected some profiles that should not be used, like in 0.1% of all encrypted communication in form of the RSA_EXPORT_WITH_RC4_40_MD5 cipher suite, which only offers a 40 bit key length or in the case of 0.01% of the encrypted communication in form of the RSA_WITH_NULL_SHA cipher suite, actually offering no encryption at all. From this information national representative can disseminate guidelines for a secure use of the Internet for agencies, companies and citizens.

7 Further analyzing of the statistical raw data

The statistical raw data we are already collecting is almost certainly concealing very interesting information for a variety of scientific disciplines. Collecting raw data from a plethora of sensors will result to a huge amount of information. Analyzing our special type of statistical raw data will require new intelligent techniques and algorithms. A very interesting discipline will be to explore these techniques and in addition to investigate, which procedures from the areas of machine learning and artificial intelligence can be adopted for our purposes. We aim to search and discover important and critical information that are not easily located by traditional methods. For example, by linking the absence or presence of parameters at certain events (using data mining for instance [Wend06]). This is just a very basic example, unfortunately our resources at the moment only allow very little and focused research. Thus, we are very interested in extending our existing cooperation in this area with interested research institutes and we can offer a large amount of past and present data as well as the sensor and evaluation technology.

The focus is not simply on technological, security or performance related information, but we also want to provide a broader multidiscipline vision by including aspects from the disciplines like economy and sociology. The driving force behind this idea is the observation that similar environments exhibit different behaviors, when using similar technologies. For example, we have observed that one of our sensors in Brazil placed at a university records different usage patterns than our sensors at German universities [Ricc08]. We believe this different usage of technology might have to do with sociological aspects and performing further research in this direction could provide stunning results. These results could for instance be very beneficial for companies looking for early adopters to test a new technology.

8 Global View

As illustrated, the Internet Analysis System can be used to generate a local view of IP-based networks. These networks, ranging besides others from companies', Internet Service Providers', Content Providers', Universities' networks, come along with very different characteristics, like obviously the total number of packets passing the sensor. The local view already helps the operator to monitor the network, but a global view is even a lot more valuable [Proe05][Tsch08]. It can be used to compare the local situation with an authentic global view to detect abnormalities, which might help to confirm or dismiss the detection of local attacks or events. The global view is valuable to a number of other relevant stakeholders as well, like for national assessment centers.

To generate this global view partners are invited to join by frequently sending a summary of their local view to a centralized evaluation system (Fig. 8). From this authentic data the common global view is generated and transferred back to the participating partners. Due to the structure of the internet, this can only be accomplished with the support of the partners. So far nobody can offer this kind of global view, so you cannot just buy it someplace.

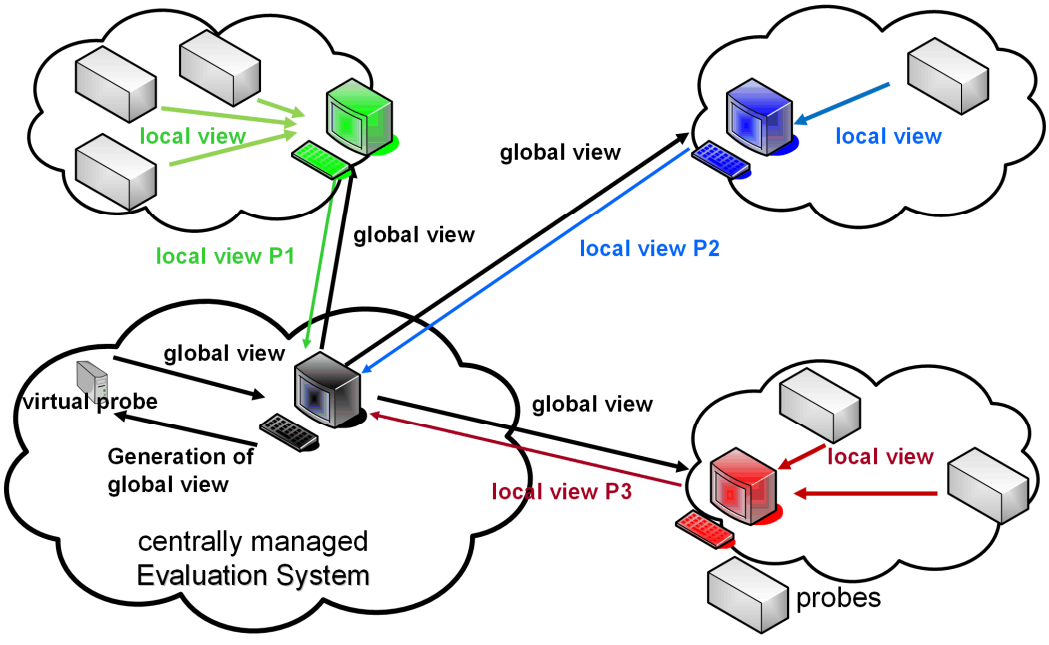

Fig. 8: collaboration for a global view

In Fig. 9 an example of the possible confirmation of an attack with malware attached to e-mails by the use of the global view is given. The local view shows our partner that the number of e-mails with a zip-attachment has abruptly increased, which is an indication for an abnormality and possibly for an attack. To verify whether this is a local phenomenon, which would be speaking for a directed attack towards the local partner, the event can be compared to the global view. Doing this, we can find out that only a few partners have recorded this phenomenon at this specific time. Therefore, the event could be a directed attack against a selected group, for example banks or insurance companies. Other partners, which have no problem at this particular moment, can use this information to protect their organizations in advance. The centrally management Evaluation System will also be able to detect cyber war activities.

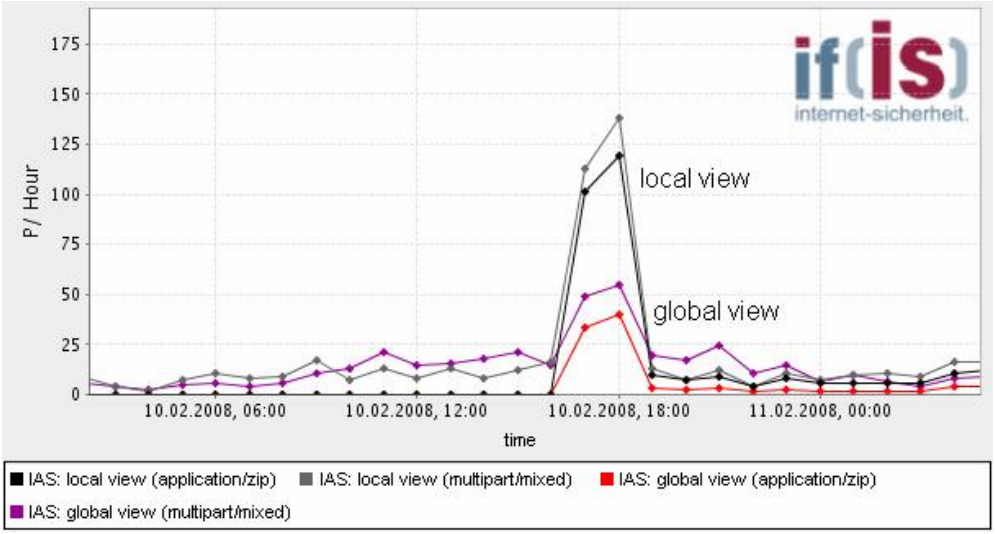

Fig. 9: Example of malware detection

A further challenge will be to deal with the different natures of networks, that can be monitored with the sensor technology. The traffic passing the sensor is characteristic for the types of services, which are provided by this different kind of network. Therefore, a content provider has a different profile of traffic as a university. To improve the outcome of the global view for these partners, they could be grouped in logical units. Each group can have their individual global view, which can be combined to the common global view of the Internet. The challenge is that not all networks can be grouped that easily, due to their diversity.

In addition, we have to deal with different time zones and inconsequent daylight saving time regulations all over the world. We cannot just agree on a global time for the distributed system of sensors, because the local time reflects in the usage of services and therefore in the raw data. The traffic passing the sensor consists roughly of human and machine initiated communication. The human part is highly affected by the different time zones, as people tend to sleep during night and work during day time. The human part is also affected by cultural issues (sociological aspects) like the long lunch break in Spain (Siesta), due to the hotter climate. Besides the fact that global attacks initiated in the US by a human, might strike in Europe at night time, we also have to face the problem that no existing real time sensor can

decide for sure by analyzing the passing the communication data, if the connection was originally initiated by a human or a machine. It is not like we could build on the work of Turing, trying to test whether the communication partner is a machine or a human. The sensor is only monitoring passing traffic in real time, which in a lot of cases are only fragments of packets of the different applications.

If a segmentation of the communication into human and machine initiated would be possible, we could introduce a global time for all machine initiated communication worldwide and consider the local time for the human initiated communication. Sometimes the communication parameters give an indication, because for example some protocols are by specification for machine to machine communication. But most protocols are shared in use or might not be used in correspondence to the specification or RFC. On top of this, we have to acknowledge that the machine initiated communication is by part also influenced by the local time, since most routine jobs are run at night, to reduce the traffic load on the network. At this point further analysis is necessary. It could for instance be that looking at a global perspective the machine initiated portion perishes in noise.

We have a strong feeling based on the results of a diploma thesis [Ricc08] that a limited but well chosen selection of parameters might already be sufficient to compare the local situation with the global overview to detect possible events. This is the great advantage of this system colleting statistical raw data, enabling the utilization of findings of this mathematical discipline. In the long run we can further extent the selected parameters, if this should become necessary. At the moment we are working on selecting and roughly splitting up the parameters. From there we can build a global view, which considers the relevant time changes for each parameter of each sensor.

9 Forecasts

The researches have developed some extensions for the IAS, one of which allows to make certain long-term and short-term forecasts. The long-term forecast can be conducted with or without seasonal impacts. One finding is that the trend of the use of an technology is clearly noticeable without considering seasonal impact. The use of “linear regression” offers the greatest accuracy for long-term forecasts. In the case that there is some heavy noise in the data methods of smoothing are more precise.

In the area of short-term forecasts the consideration of seasonal aspects is very important for instance day- and night changes, lunch breaks, working day and holidays. Our current findings show that when you break down to an interval of hours one should prefer the method of “linear regression”. The shorter the interval the more exact the “Holts-Winters” method turns out to be, which should be used for intervals of minutes. All findings are available in German and can be found in the referred document [Deml08a].

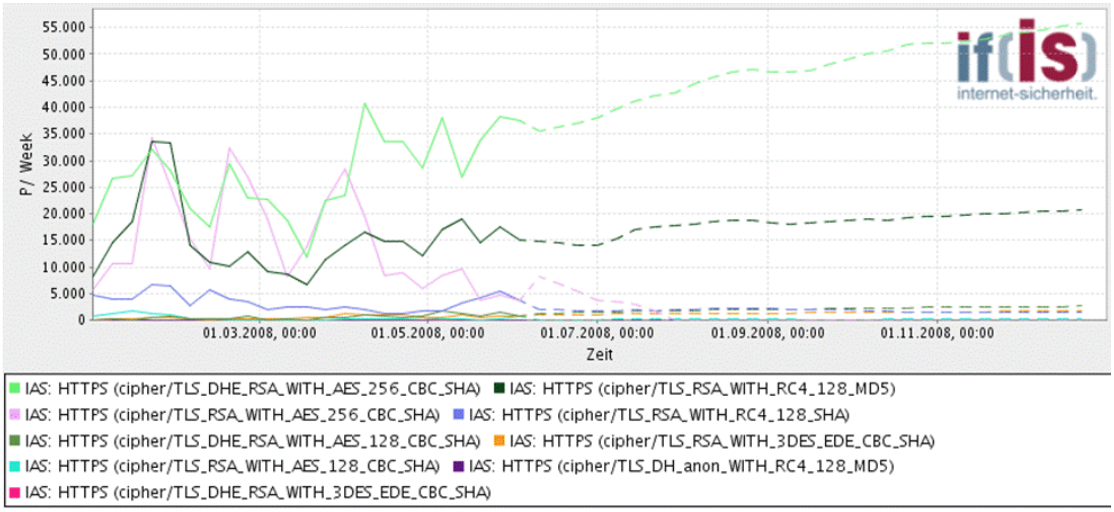

Fig. 10: forecast of TLS cipher suites in use

10 Conclusion

The provided information is supposed to help you understand, how the Internet Analysis System works and what the Institute of Internet Security has conducted so far in this area. This should help you determine, whether the system and our visions are interesting to you and your organization's proceedings. We are looking for partners in various roles ranging from sensor partners using our sensor technology to monitor their administrative domain to new research partners from various scientific disciplines. As a partner you will have full access to our technology and collected raw data free of charge.

We hope that this White Paper has aroused your interest. Please do not hesitate to contact us:

Malte Hesse, <hesse[at]internet-sicherheit[dot]de>

Prof. Norbert Pohlmann, <pohlmann[at]internet-sicherheit[dot]de>

References

[Deml07] Mathias Deml, IAS Rohdaten Transfer System (internet analysis system raw data transfer system), University of Applied Sciences Gelsenkirchen, 2007.

[Deml08a] Mathias Deml, Evaluierung von Prognoseverfahren für das Internet-Analyse-System (evaluation of forcasting methods for the Internet Analysis System), University of Applied Sciences Gelsenkirchen, 2008.

[Deml08b] Mathias Deml, Aufbau eines Signaturerkennungssystems für die Statistik-Sonden des Internet-Analyse-Systems, (development of a detection system for patterns for the statistical probe of the Internet Analysis System) Master Thesis, University of Applied Sciences Gelsenkirchen, 2008.

[Dier06] Stefan Dierichs: Eine strukturelle Analyse des Internets (a structural analysis of the internet), Diploma Thesis, University of Applied Sciences Gelsenkirchen, 2006.

[DiPo05] S. Dierichs, N. Pohlmann: "Netz-Deutschland" (German Internet Infrastructure), iX - Magazin für professionelle Informationstechnik, Heise-Verlag, 12/2005.

[Hees06] Uwe van Heesch: Entwicklung eines Plugin basierten Analyse-Frameworks für das Internet-Analyse-System (development of a plugin-based analyzing framework for the Internet Analysis System), Diploma Thesis, University of Applied Sciences Gelsenkirchen, 2006.

[Kort06] Stefan Korte: Internet-Frühwarnsysteme (internet early warning systems), Diploma Thesis, University of Applied Sciences Gelsenkirchen, 2006.

[Oster06] Thomas Ostermann: Internet-Verfügbarkeitssystem (internet availability system), Diploma Thesis, University of Applied Sciences Gelsenkirchen, 2006.

[Pohl07] N. Pohlmann: “Probe-based Internet Early Warning System”, ENISA Quarterly Vol. 3, No. 1, Jan-Mar 2007

[Pohl05] N. Pohlmann: "Internetstatistik" (statistics of the internet), Proceedings of CIP Europe Publisher, B.M. Hämmerli, 2005.

[Proe05] Marcus Proest: Die globale Sicht auf das Internet (the global view of the internet), University of Applied Sciences Gelsenkirchen, 2005.

[Proe05a] Marcus Proest: Entwicklung einer Sonde für ein Internet-Analyse System (development of a sensor for an internet analysis system), Diploma Thesis, University of Applied Sciences Gelsenkirchen, 2005.

[Ricci08] Gianfranco Ricci, Betrachtung der vom IAS gesammelten Kommunikationsparameter auf Relevanz zur Anomalie und Angriffserkennung (evaluation of the relevance for the detection of abnormalities and attacks of the communication parameters collected by the internet analysis system), Diploma Thesis, University of Applied Sciences Gelsenkirchen, 2008.

[Tsch08] Sven Tschöltsch, Konzeption und Realisierung einer globalen Sichtweise auf das Internet zur Bewertung der eigenen Sicherheit (concept and realization of a global view of the internet for a better evaluation of the local security situation), Diploma Thesis, University of Applied Sciences Gelsenkirchen, 2008.

[Wend06] Svenja Wendler: Entwicklung eines Anaylsemoduls zum Internet-Analyse-System – Finden von Strukturen im Internetverkehr in Form von Assoziationsregeln (development of an analyzing module for the internet analysis system – discovery of structures in the internet traffic consisting of association rules), Diploma Thesis, University of Applied Sciences Gelsenkirchen, 2006.