

# Internet Analysis System

Prof. Dr. Norbert Pohlmann · Marcus Proest

Institute for Internet Security  
University of Applied Sciences Gelsenkirchen  
Neidenburger Str. 43, D - 45877 Gelsenkirchen

{norbert.pohlmann | marcus.proest}@internet-sicherheit.de

## Abstract

We have all experienced the situation: you are sitting in a traffic jam and all you can see is a long line of cars in front of and behind you. In this situation, without any assistance, you do not have an overview of the problem. There is no direct information concerning why the traffic jam has come about, how long it is, at what point of the traffic jam you are located or - the most important information - when the traffic jam will be over. As this is a problem faced on a day-to-day basis by thousands of motorists, solutions have been developed to overcome the lack of information. In Germany there is a close network of traffic counter loops which record the traffic volume and situation on the motorways/freeways. Important information about traffic jams is provided by means of radio announcements, SMS, telephone and the Internet, while modern navigation systems process the information directly when planning the route to be taken. Through the use of these resources, motorists are "liberated" from their constricted local view of the situation and can take decisions in good time on the basis of the global information available, e.g. leaving by the next exit and using an alternative route.

This situation can also be applied to the perspective that the network operators have today of the Internet. As a rule they have only a local perspective, i.e. an overview of their own network segments and the communication data that are transferred. If problems occur here and are detected, they can be rectified quickly and systematically. However, if it becomes apparent that a problem has occurred that is not within their own domain of action, or if the required perspective is lacking, the situation is more difficult. It is often not clear where the problem comes from, and for the correction of the problem we are reliant on third parties.

The global view required in order to detect the problem and select the correct solutions is missing. Such a global perception is difficult to achieve on the Internet as people like to play their cards close to their chest. The precise internal network structure, communication connections and topologies are often treated confidentially by the network operators.

Furthermore, in order to achieve a global perspective, there are a few challenges that have to be solved: communication data are relevant in principle to data protection, the quantities of data are enormous, the datarates are sometimes so large that they cannot always be analyzed "live", while long-term storage of the communication data in order to observe long-term developments appears to be impossible. Moreover, the question also arises of who feels responsible for creating a global perspective? Nevertheless, the Internet has developed into an omnipresent medium over the past few years, without which very large areas of the economy, research and private life would be unimaginable today. For this reason the analysis and knowledge of the medium known as the Internet in its totality is of particular significance in order to be able to assess its development and guarantee the future functioning of all the services it provides.

The constantly growing importance of the Internet for our knowledge and information society makes it necessary to analyze and be acquainted with its status beyond the limits of the individual network op-

erators. Only precise knowledge of the normal status makes it possible to detect anomalies which influence the functionality of the Internet.

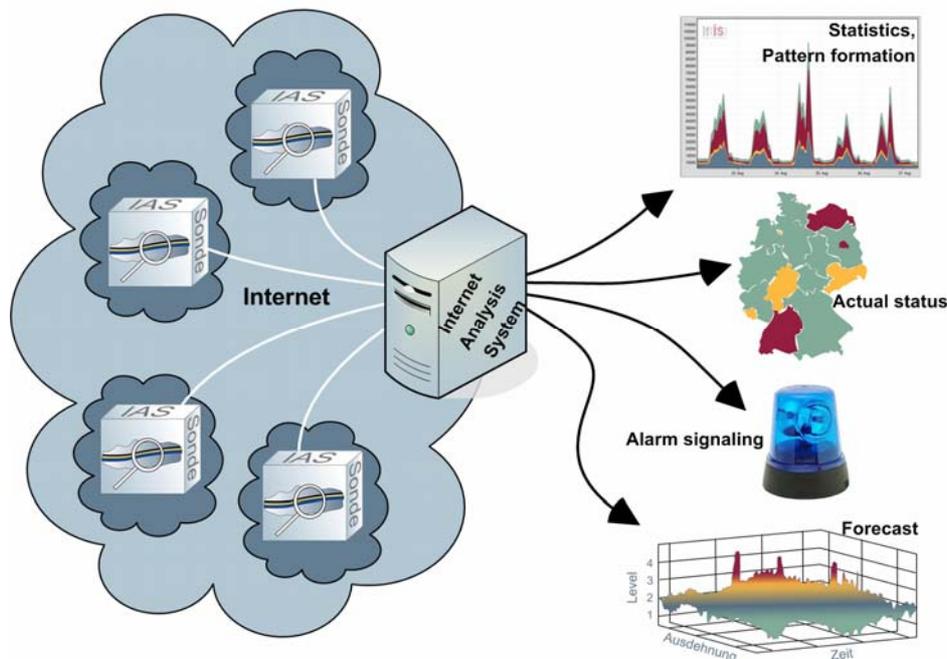
With the help of the probe-based Internet Analysis System, which is currently being implemented as a research and development project of the Institute for Internet Security at the University of Applied Sciences in Gelsenkirchen in collaboration with the German Federal Office for Information Security (BSI), it is intended to create and analyze local and above all global perspectives in order to make the generation of early warnings possible.

Particular focal points of the project are the collection of information in compliance with data protection regulations and optimization of the amount of information, so as to be able to store information in the long term and therefore allow the analysis of trends and developments over long periods.

## 1 Aims and Task of the Internet Analysis System

The task of the Internet Analysis System on the one hand is to analyze local communication data in defined subnetworks of the Internet, and on the other to create a global perspective of the Internet by bringing together the large number of local perspectives.

The functions of the Internet Analysis System can be divided up into the four subsegments of pattern formation, description of the actual status, alarm signaling and forecasting.



**Illustration 1:** Tasks of the Internet Analysis System

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. On the basis of this knowledge a search is carried out for anomalies among the current measured values and the causes of status changes 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. .

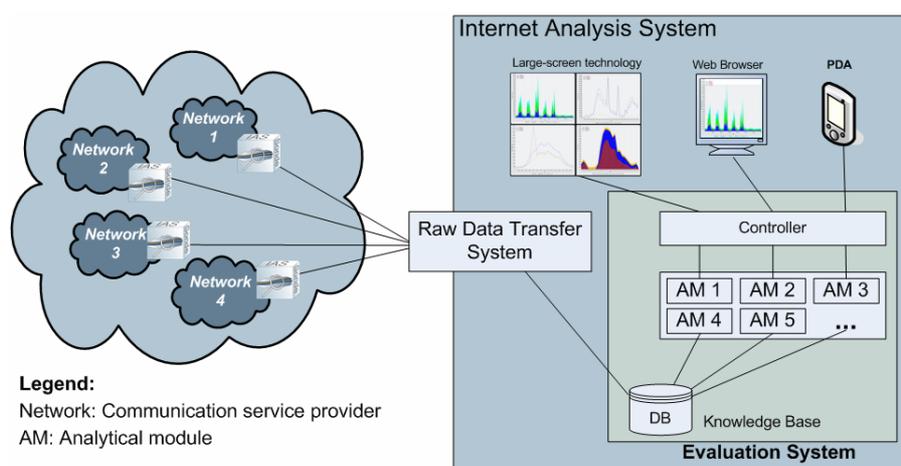
With knowledge of the current status of a communication line and the use of historical - i.e. previously collected - information (knowledge base) it is possible in the case of significant changes to traffic volumes or communication data to generate a warning message, on the basis of which measures can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

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.

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 Internet status changes. 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 [Pohlmann2005].

## 2 Mode of Operation of the Internet Analysis System

The Internet Analysis System consists of probes which passively access the network traffic of the communication lines of various networks and count communication parameters at various communication levels. In an evaluation system the communication parameters are evaluated from various aspects and displayed in a clear manner. Illustration 2 shows the interrelationships between the components involved in the Internet Analysis System.



**Illustration 2:** Components of the Internet Analysis System

In order for the Internet Analysis System to be able to supply meaningful results it requires as great a quantity of raw data as possible, i.e. very many counters of various communication parameters at all communication levels over time. All the analyses performed by the evaluation system are based on these raw data. They consist of aggregated counter readings with which the system is provided by various networks. The left-hand side of illustration 2 depicts the Internet, which consists of a combination of numerous networks. In each of these networks telecommunication service providers (ISPs, companies, universities ...) assume the tasks, for example, of providing Internet access to end-users, content or other services. The Internet Analysis System acquires its raw data from probes implemented and operated by the network operators. The raw data are transmitted via the Secure Raw Data Transfer Protocol (RDTPs) specified for this purpose.

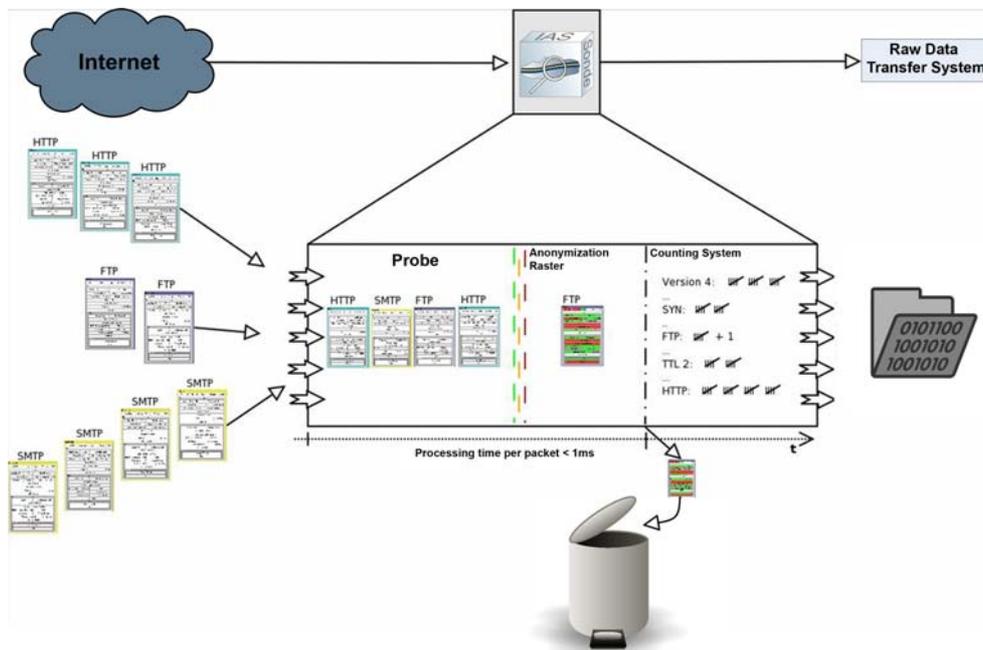
The probes can send the raw data to one or more evaluation systems. Each network is able to perform its own analyses with its evaluation system. In order to achieve a global and representative perspective of the Internet, probes have to be operated in various types of networks, such as the Global Tier One Provider, Transit Provider, Eyeball Internet Service Provider, Content Provider and Business Networks, as well as various regions (see also in Internet Germany [Dierichs2005]).

### 3 Tasks and Mode of Operation of the Probes

It is the task of the probes to extract information from a communication data stream which provides details of the status and use of the communication line and the network or networks behind them. Here all information should be retained which is required to detect misuse, a misconfiguration, trend developments or an attack situation. At the same time, however, the quantity of information should be restricted to the minimum required so that the information can also be considered and analyzed retrospectively over long periods. A further important point for the operation of the probes is that the information extracted by the probes, the so-called raw data, does not contain any information relevant to data protection law.

Technically the network connection is accessed passively and the communication parameters of the various protocols at the communication levels counted. The results of the counts are transferred at defined intervals to the raw data transfer system.

### 4 Principle of Raw Data Collection



**Illustration 3:** Principle of raw data collection

Illustration 3 shows the principle of raw data collection by the probes. This is divided up into three sections. The Internet is represented on the left. Packets of three different application sessions are shown: related HTTP packets, an FTP session and an SMTP session. The probe is located in the middle of illustration 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 law. 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. For example, in illustration 3 the accessing of the FTP packet is recorded by incrementing the FTP counter by 1. The raw data are therefore aggregates of counters, i.e. counters of communication parameters that have appeared at the various communication levels over a defined period. The packet - in illustration 3 an FTP packet - is immediately deleted physically, i.e. irreversibly and without trace, by the probe [Proest2005].

<u>ID</u>	<u>Description</u>	<u>Count</u>
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

**Illustration 4:** Counting system in the probe

Reconstitution of the context of a packet or only a communication parameter is not possible or necessary. At definable intervals the counter readings (raw data) of the probes can be transmitted to the raw data transfer system. All of this information is completely anonymous, as shown in Illustration 4. On the right after the colon are the counter readings for the header information specified on the left. Each line stands for a counter. On the left-hand side of the colon is the count-if function (appearance of the corresponding communication parameters) and on the right the number of packets which contained the communication parameter during the defined measurement period. For example, line 2 of the raw data shown indicates that 1,123,149 packets with the IP protocol number 17 (UDP) appeared in the prescribed time. The count-if functions and their codings are specified in a versioned XML file.

The raw data transfer system functions as a server to which the probes can connect in order to transmit their raw data for a defined period. This is a unidirectional connection, meaning that a connection can only be established from the probe side. A probe can transmit the raw data to one or more raw data transfer systems. An example of a typical configuration is that every 5 minutes the raw data, for example 20 kilobytes in size, are sent to its own and a central raw data transfer system.

As the raw data are only a statistical formulation of the actual communication data, it would also be sufficient if not every packet were considered, but for example only every 10<sup>th</sup> packet. This aspect can be a pragmatic solution in the case of very high communication data rates, without producing a different result from a statistical point of view.

## 5 Evaluation of the Collective Raw Data

The actual evaluation and processing of the collected information takes place in various analytical modules (AM) of the evaluation system. In Illustration 2 these are designated "AM1" - "AM5". The modules procure the information exclusively from the knowledge base (raw data and evaluation results). The aim of the various modules is the compilation of profiles, statistics and interrelationships, as well as the detection of where threshold values are exceeded and the graphic processing of the raw data and evaluation results.

As the raw data consist of completely anonymous information, they could also be exchanged between various network operators or collected at a central point in order to be used as the basis for a global perspective and analyses of an IT early warning system.

## 6 User Interface of the Internet Analysis System

There are many conceivable methods of displaying the results of the Internet Analysis System. Illustration 2 contains examples of the following: large-screen technology, a Web Client and a PDA. The large screen technology serves the purpose of continually updating the display of certain statistics, profiles and current statuses. By means of an intelligent client, more extensive analyses can be carried out with the evaluation system and the results recorded. Additionally, warning messages from the system can be received in mobile form, for example by a PDA, so that an initial overview of the hazard situation can be obtained. Currently the Internet Analysis System uses a stand-alone client as a front-end. Through this client all counter readings can be displayed for freely selectable periods. It is possible to include further functions by means of a plug-in system.

## 7 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 possibilities of the current status of the Internet Analysis System. At present there are approximately 300,000 different counters of communication parameters incorporated for the various communication levels. This large number clearly shows how complex the results can be.

### 7.1 Transport Protocol Distribution

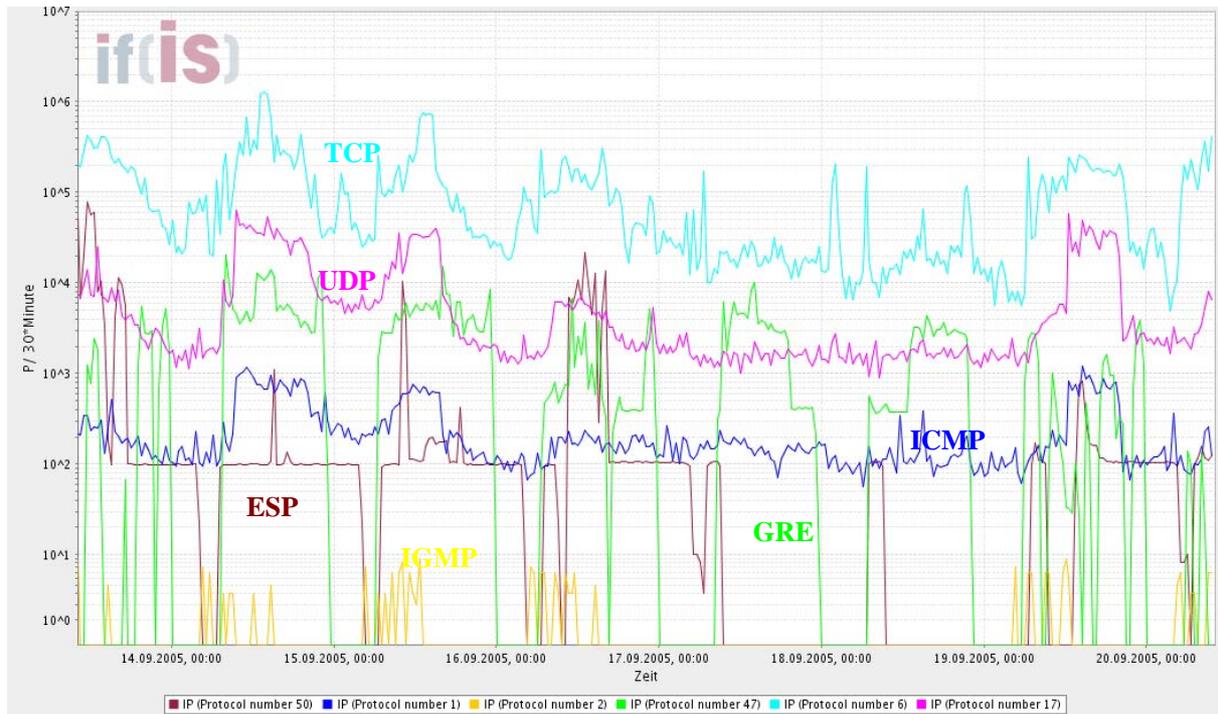
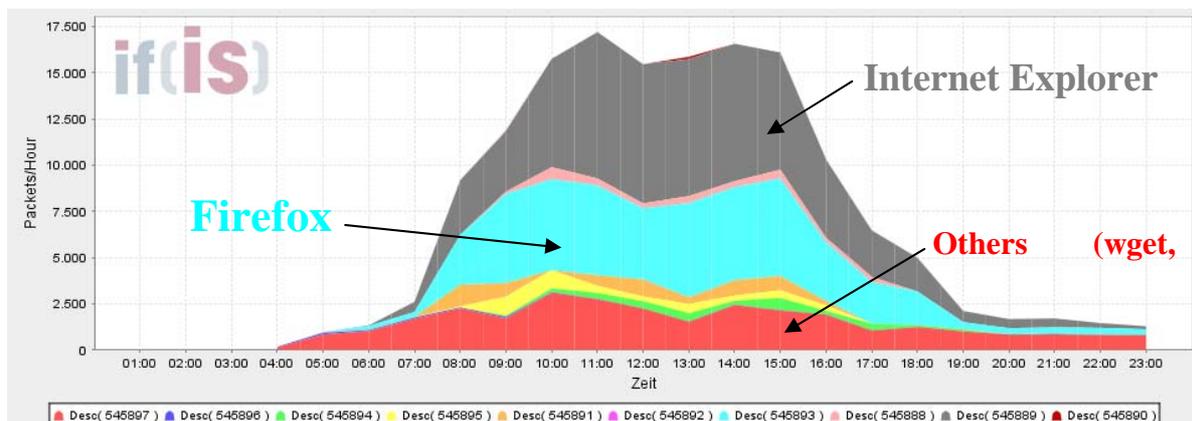


Illustration 5: Protocols of the transport layer

Illustration 5 shows the distribution of the protocols of the transport layer used over a period of several days for a specific communication line. From the past the Internet Analysis System knows the profile, the standard deviation and from this can 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 and DNS.

### 7.2 Browser Distribution (Technology Trend)



### Illustration 6: Browser distribution over time

Illustration 6 shows the distribution of various browsers over a period of one day for a specific communication line.

Here we can see the daily profile of the various browsers. The difference between manual use (e.g. Internet Explorer and Firefox) and automatic use (e.g. wget) over the course of the day can be clearly seen.

It is noticeable that these statistics in no way refer to a Web server, as is usual today. In contrast, these statistics refer to a communication connection.

## 7.3 Types of E-mail Messages

Illustration 7 shows the ability of the system to record the statistics of the headers of the e-mails sent by SMTP. The distribution can provide information on general communication behavior, as well as deviations from it. Illustration 7 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, text/html and multipart/alternative content types. As a rule, e-mails with attachments are provided with the multipart/mixed content type. A mixed form is e-mails with the multipart/related content type. Here, for example, images are integrated directly into the text. If these e-mails are included in the e-mails with 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 content type. An abrupt change to 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.

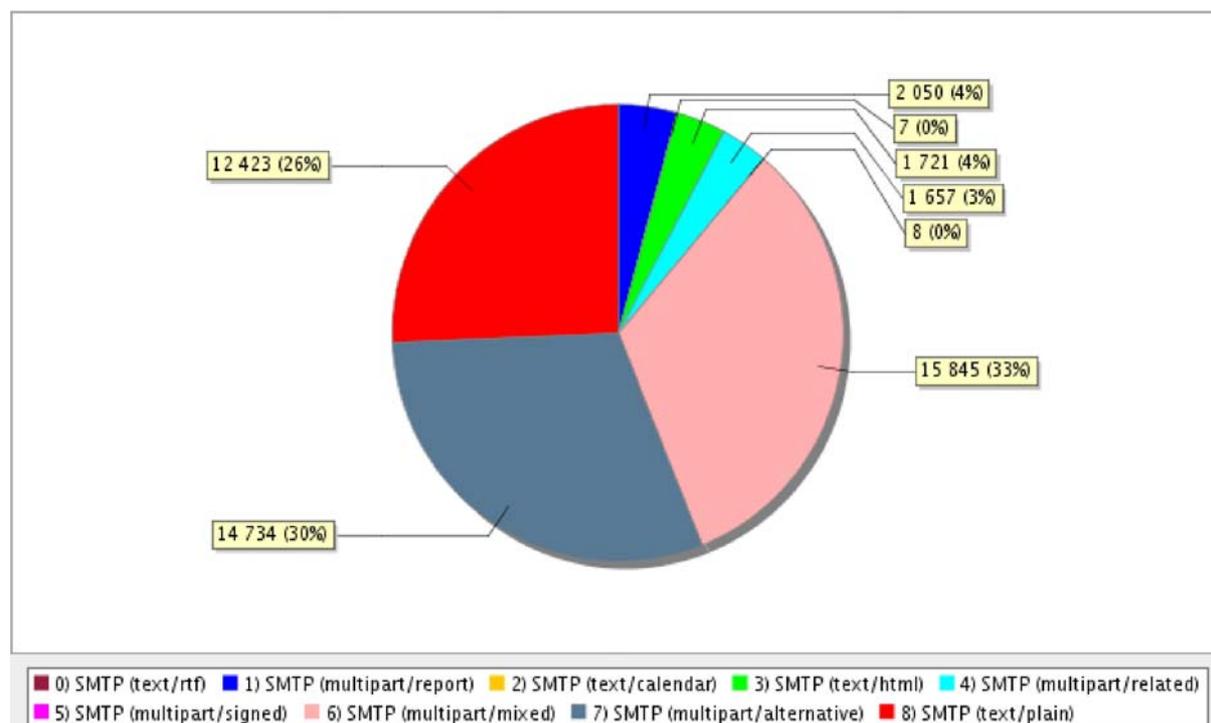


Illustration 7: Distribution of e-mail "Content Types"

## 8 Uses of the Internet Analysis System

Special operators with their own evaluation system are informed by the system when problems occur and can use the wide-ranging resources offered by the system in the analysis of the problems. A clear representation of the current status of their own system and an informative reporting system represent the basis for reliable network operation.

If their own system forms part of a central evaluation system, information on problems with a participant is passed on in good time, so that countermeasures can be taken at an early stage by all.

## 9 Perspectives

With the help of the probe-based Internet Analysis System it is possible to continuously acquire raw data which provide a statistical reflection of Internet traffic. Through the evaluation of the raw data at the various communication levels, such as the network layer, transport layer and application layer, it is possible to derive very detailed information.

By analyzing the results of various probes it is possible to depict a global perspective of the Internet and define warning levels in the case of problems, such as infrastructure failures or attacks. Further analyses of the raw data allow the forecasting of trends in the use of protocols, network services and attacks.

A further important process is the exact analysis of the raw data in order to correlate the results with information from other network analysis tools. With the interrelationships that are created it is possible to detect patterns and thereby compile results in the form of statistics, patterns, attack scenarios and, as a subject of research at the moment, forecasts. In turn, these results can be incorporated into the review and evaluation process in order to acquire ever more accurate results at an increasingly earlier stage. Here the extent of the raw data pool is of crucial importance: the more raw data that can be analyzed - with respect not only to the number of various probes and their positioning, but also the time period of the information - the more accurate results are to be expected. This applies in particular to the forecast analyses, as algorithms which are suitable for carrying out the corresponding analyses require a particularly large quantity of initial data.

The Institute for Internet Security has successfully completed the second development phase in collaboration with the BSI. The system has proved its worth during operation. Partners are now being sought to join the network and operate further probes and evaluation systems. Organizations who wish to operate only one probe will receive an informative report about their network traffic once a month from the central evaluation system.

In addition to operation of the Internet Analysis System, the project will also be moved forward to a third development and research phase. In order to obtain additional information, further protocols will be implemented and existing ones extended. Furthermore, the evaluation modules and attack detection will be expanded.

In the field of research, statistical methods and data mining algorithms will be used to design intelligent processes intended to accelerate the evaluation process. The quality of the results is steadily improving as a result of the continuous analysis and the incorporation of the experience gained of the analytical process.

By its very nature the Internet is extremely flexible and complex and subject to daily changes. In order to be able to provide statistics on the Internet and for any future technologies, the Internet Analysis System also has to be extremely flexible in its basic form.

For an extensive IT early warning system the Internet Analysis System can, for example, be expanded by an availability system and log data evaluation system.

It is therefore important that this system is used both flexibly and globally in order to allow - just as in the traffic jam on the motorway/freeway - a global perspective of the structures and dangers.

Further information:

Institute for Internet Security, <https://www.internet-sicherheit.de>

Federal Office for Information Security (BSI), <http://www.bsi.de/english/index.htm>

## References

[Dierichs2005] Dierichs, Stefan; Pohlmann, Norbert: "Netz-Deutschland", iX - Magazin für professionelle Informationstechnik, Heise-Verlag, 12/2005

[Pohlmann2005] Pohlmann, Norbert: "Internetstatistik", Proceedings of CIP Europe 2005, publisher: B.M. Hämmerli, S.D. Wolthusen; Gesellschaft für Informatik, Bonn 2005

[Proest2005] Proest, Marcus: "Internet-Analyse - Ein Blick in die Dunkelheit", conference: Internet security 2005; <http://www.internet-sicherheit.de/center-berichte.html>

## Keywords

Internet Early Warning System, probe-based Internet Analysis System, global view, internet perception, internet monitoring, meshed internet analysis, internet statistics