Nontechnical SPAM Detection Paradigm in Unified Communications Systems

Moritz Giesecke
School of Engineering, Pforzheim University of Applied Sciences
D-75175 Pforzheim, Germany
moritz.giesecke@hs-pforzheim.de

***Abstract*—The recognition and filtering out of unwanted messages in technical communications media presents an ever more difficult challenge. The best-known of these problems is with ubiquitous e-mail. Most e-mail sent are unwanted spams. In order to protect the recipient the most diverse applications must be used. Longer observations have shown that spam is continually adapted and is able to overcome the most up-to-date recognition programs. In the future the most widely different communication methods are growing together such as e-mail, telephony and others, so that soon we will be able to speak of unified communication. There is a danger that these other communications media will increasingly become the target of new types of spam. On the other hand this logical union opens up new possibilities for spam recognition. In this paper, a behaviour-based evaluation paradigm is introduced which works on a uniform basis for all communications media. It uses an evaluation of the three parameters of abstracted times of usage, distance of communication partners and costs. All communication events between media using actors create a social network whereby the actors are clustered according to their social proximity. The evaluation of spam is a result of the actors and cluster specific communication behaviour up to a point. In this way a new non-technical level of analysis is created, which spammers can only overcome with difficulty. Likewise the problem of limited focus in network centred filtering programs is dealt with. The presented filtering paradigm can be used unitary in all technical communications media and works with the same three nontechnical parameters at a behavior-based level.**

***Index Terms*—spam, spit, unified communications and social networks.**

I. Introduction

In modern communications media the proportion of unwanted messages is continually growing. A classical example is e-mail spam, which has appeared since the widespread use of e-mail services. Normally, these are differentiated between unsolicited commercial mail (UCE) and unsolicited bulk e-mail (UBE). Both forms are normally characterized as spam [1]. The particular societal and economic meaning of this amount, around 120 billion spam e-mails per day, or calculated at up to 20 spams per day per person is fatal [2]. Typical return rates are under 1 per thousand, depending upon the quality of the spam [3]. By processing these spams economic damages are incurred in the form of lost working time, server and energy costs as well as the irritation of the users of the e-mail service. By today, classical spam technologies are no longer used only for advertising purposes; they are used for fraud, typically called phishing [2].

Until now various classes of processes were used, based upon the individual e-mail infrastructure or the users' mail boxes to protect against incoming spam. These differ according to granularity, effectiveness as well as complexity of the filtering. In a typical mail server, these filters are arranged in a cascade, see Fig. 1. Starting with a Firewall, all incoming connections from IP addresses recognized as known spam senders are blocked. After this first IP address based list process, different black and white lists with known spam servers and e-mail addresses are queried. These come from specialized companies, which put a lot of effort into the finding of the most up to date and correct data. Typically, the highest level of effectiveness is achieved through recognition of spam e-mails and through the avoidance of false positives. For this, three different lists are used and then evaluated with a two out of three decision. Afterwards, information in the e-mail header is checked to see if the address name and server details are correct for the domain to send e-mails by searching DNS records, which contains appropriate mail exchange information [4]. After this, a specification of the SMTP protocol is exploited using the grey listing process [5]. In doing so, a temporary problem in receiving the e-mail is simulated in the users own server. Real mail servers wait a certain amount of time and try a new delivery; typical spam senders on the other hand drop the repeated delivery of the spam e-mail. A further recognition method, which consists of comparing as many as possible of received e-mails is the well known process called *Distributed Checksum Clearinghouse* (DCC) with a distributed checksum filter [6]. This consists of cross server boundary checking for the existence of the same or similar e-mails. Following this, e-mails are evaluated according to content based on the signalling process method. Typically, Bayes-filters are used, which must firstly be trained with the typical appearance of spam and reasonable e-mails [7]. Afterwards, on the basis of this training, the e-mails were scanned for these patterns of learned words and then weighed against each other. This results in a decission about the current received e-mail, whether it contains spam or not. In addition to these generally easily accessible procedures, as they are reproducible from open source software, there are different commercial service providers with proprietary procedures offering the evaluation of incoming e-mails. An example named here is Cisco SenderBase, which works off of a central database containing reputation values of individual e-mail addresses and organisations [8]. The evaluation process uses more than a hundred different parameters for the evaluation [9]. An important critical point is the central capture of the e-mail traffic occurs nearly in real time and the long term storage of the same. Parallel to these technical filtering processes, the introduction of a global legal barrier for allowed e-mail marketing

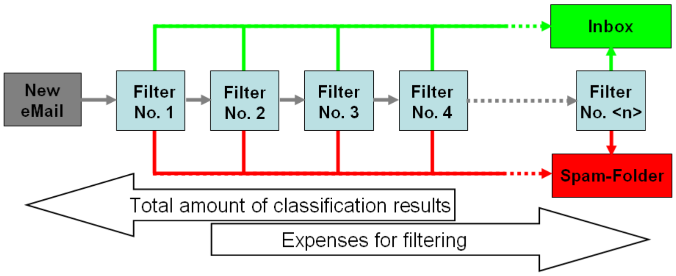

Figure 1. Spam filter cascade with cost-benefit relation

ISBN: 978-1-61208-105-2

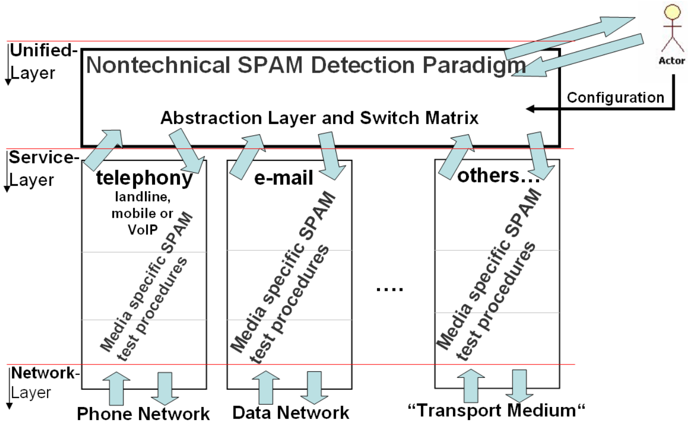
Unified-Layer
Nontechnical SPAM Detection Paradigm
Actor
Configuration
Abstraction Layer and Switch Matrix
Service-Layer
telephony
landline, mobile or VoIP
e-mail
others…
Media specific SPAM test procedures
Media specific SPAM test procedures
….
Media specific SPAM test procedures
Network-Layer
Phone Network
Data Network
“Transport Medium“

Figure 2. Unified Communications System in an abstract view

since at least the end of the 1990s took place, i.e., in the USA and Germany USA and Germany [10] [11]. In spite of a few, but deterring judgements there has been no visible reduction in spam. The leading suspects are known in part by name and with a photograph [12].

A. Characteristics and transformation of spam

In order to be interesting for the spam-senders, a few characteristics must be present in the communication media. Mainly, low cost, so that an individual spam process can not incur any costs. Because of the low rate of reply gigantic amounts of individual spams are sent. Equally important is the possibility of sending a *variable content*. In order to provoke a response from the recipient of the spam, the spam appearance must be varied. *Limited traceability* meaning that the spammer tries to conceal their true identity to avoid trouble with the recipient of spam. There can be possible civil suits for damages and financial compensation and severe criminal consequences. *Simple completion*; meaning that the recipient of the spam should easily be able to respond to the spam. Typically, spam has a feedback link, which logically lies as close as possible to the communication media of the spam. For example, a successful spam e-mail pulls the user who received it directly to a web page, which may instantly be opened with one mouse click.

In spite of the laws against spam and other legal instruments available there is little help on the way towards a more tightly ordered e-mail framework. That is why new technological evolution must be continuously developed and implemented in order to act against the continuous flood of spam. Communications media are the target of spam as soon as the above mentioned characteristics are fulfilled. In addition to the old e-mail spam problem, the spam over internet telephony (SPIT) is growing recently, but is still not as intensive as traditional e-mail spam. With a telephone call, the recipient is provoked into giving a reaction which allows the spammer to make a profit. An example hit German customers that used a SIP-based VoIP connection in the first mass spamming in September 2008 [13] [14].

B. The idea of Unified Communication

The users of modern communication technologies are taken in more and more by the complexity of the technology and the operating effort for the user of different communications media such as e-mail, telephony (land lines, mobile or cell phones and voice mail services) as well as multiple specialized services (Instant Messaging, Pager, Groupware solutions). In addition to various user variations and configuration options on the part of the recipient's end it is also difficult for the communications initiator to reach the desired communications partner with the correct communications medium. The origins lie in the asynchronous communications media such as e-mail, for which here the term Unified Messaging is used. The idea of Unified Communications (UC), is the managing of different types of communications media bundled in one location with a supporting function to relieve the user. It is not totally clear in the general language usage as of when Unified Communication can be spoken of or rather, which criteria must be completely fulfilled. For the consideration of this paper it is assumed that in addition to the other aspects of UC technology a central instance exists that captures all communication procedures, manages all of the user's preferences and can appropriately influence the communication. As an example calls can be rerouted to a voice mailbox and the message can be sent by or as an e-mail. It is equally possible to think of calling up e-mails by telephone and having them read aloud using a text2speech system. The required settings would be in put by the user themselves, and thereby be part of the user's preferences. Such an Unified Communication System (UCS) scheme is shown in Fig. 2. All communications media in the UCS, are connected to the rest of the world by their different transport media in the *network layer*. Typically, this is an IP-based network for e-mail and VoIP or traditional phone networks often referred to as PSTN. Any *other* communication media are of course also conceivable. The different communications media resides, in the so-called *service layer*, where they are considered as separately existent. The usual *media specific Spam test procedures* (e.g. signalling and content evaluation), are applied here. The *unified layer* is aware of all communication events sent or received by the actor, indicated by the light-blue arrows. Furthermore, this layer makes available also all the benefit features described with the idea of UC, once it gets configured by the actor. This is the instance is the place, where the subsequent described nontechnical Spam detection is carried out.

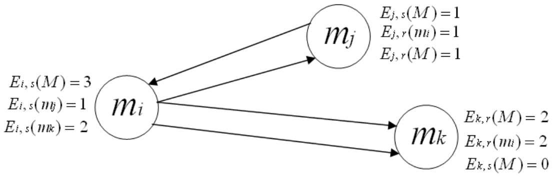

Figure 3. Social network as a graph with actor properties

C. Information gain with social networks

With the idea of the social network connections between people can be formally modelled, whereby the interactions of people can be graphed. Fig. 3 depicts the connections E, shown as edges, between the individual human actors $M = \{m_1, m_2, \dots\}$, shown as nodes. The edges arise from the performed communication events within the group of observed actors. This concept uses a communication process, for example a delivered e-mail or a finished telephone conversation or any other discrete event using any other possible communications method. In this way any interaction between people through the use of communications media can be represented. At first invisible information content is made up of exposing the relationship, the organizational structures, work processes and the influences of events. The sociological and mathematical formulated questions of the *social network analysis* (SNA) have been researched enough and have found practical applications in sociology, economics and criminalistics. Further applicable methods such as data mining can be derived from general sources [15] [16].

In large UC systems of telecom companies are I different human actors present, who have access to K different communications media. Between the system members and the system non-members from outside user communications processes take place, which can be transferred into a social network. At the observation starting point there are already a number of communications processes available so that an adequate connection density amongst the actors of the social network exists. Because of the freedom of the modern communications media, boundless interactions between them are possible and the evaluation of the social proximity network follows through the intensity of the incidences of communications events. Therefore the choice of the communications medium K is irrelevant. Important is the individual communications event E only. Thus a first approximation of the social proximity N_{soc} between two actors is given through the amount of events between them. Here is a summary of reciprocally received events by recipients (**r**$eceived$) where the syntax $E_{<acteur>,r}(<partneracteur>)$ is used:

$$N_{soc}(m_i, m_j) = E_{i,r}(m_j) + E_{j,r}(m_i)$$

A proposition with this first assumption does not take into account the relativity of the amount of communications events that the actors or partner actors sent (indicated by s) to the other actors within the observed social network. In order to be able to capture these relatively important reciprocal events, both actors are introduced with an additional proportionality factor $\frac{E_{<Acteur>,s<PartnerActor>}}{E_{<Actor>,s}(M)}$. In the sum of all E_s of an actor to a partner actor and the sum of all E_s from this actor to all other actors M, the attractiveness of the respective partner actor is determined from the ratio between the count E_s to the respective partner actor and the count of E_s to all others actors in M. Because in a UCS the human actors continuously communicate, the form of social network is seen as variable and therefore also those with the equivalent (1), determined value for N_{soc} between two actors.

$$N_{soc}(m_i, m_j) = E_{i,r}(m_j) * \frac{E_{i,s}(m_j)}{E_{i,s}(M)} + E_{j,r}(m_i) * \frac{E_{j,s}(m_i)}{E_{j,s}(M)} \quad (1)$$

That is why it is recommended to define an observation interval Δt, within which a calculation of N_{soc} (approximation) is seen as valid and must be newly recalculated. These order of magnitudes of the observation intervals result in the emergence of new E, the system performance capability towards the eradication of filtering cycles (see Section II) as well as the volatility of the current spam in comparison to the recognition capability of the filter systems.

For N_{soc} the valid conditions are that the value of the result is non-dimensional, $N_{soc}(m_i, m_j) = N_{soc}(m_j, m_i)$ and $N_{soc} \geq 0$. The clamping of such a social network can first take place after a initial observation time, meanwhile the actors have produced a certain amount of communications events E. For the new evaluation process of this paper the quality of the social network depends upon as many as possible of the intended communications processes in the network derived by means of the intended prototypical performance. Only a few spams which were able to overcome the previous filters can be tolerated as they will be detected as inappropriate.

D. Related works

Most spam recognition processes are based on the technical signalling information of the various communications media, e.g. within the transmission of e-mail typically on the level of Internet Protocol and SMTP. A further class of processes work on the basis of content, in the e-mail service as an example Bayes and Markov filters or DCC as well as VoIP methods for the differentiation of humans and machines. Furthermore there are ideas for the use of processes out of the field of social network analysis (SNA) for the recognition of spam [17] [18].

Typically these approaches are used on ordinary e-Mail traffic and use the results of various metrics to scan for the characteristics of spam. Here, primarily two general classes of procedures are widely used. On the one hand it is attempted to assign each communication participant a reputation value based on experience over a long period of time. On the other hand, the behaviour of communication participants vis-à-vis other participants can conclude the likelihood of spam.

Nevertheless, in the technical reality there exists the problem of a limited focus. Every filter instance of any such SNA based process can only work with the information that runs through the communications system used. This leads to spammers from outside the SNA based filter system, in certain instances, exhibiting no complete characteristics of spam if they, for example, only send a small amount of spam in the focus of the SNA filter systems [19]. Thereby the spammers wouldn't, in certain circumstances, be recognized as such. From the main countries of origin spam is distributed globally and it can be assumed that these will not be completely hit by the individual SNA based filter systems.

The filter processes presented in this paper primarily observe only individual system communication participants using all available communications media. It evaluates, using the method of normal user behaviour as though these successful incoming communications processes are desired. In the following, a non-technical level of analysis that circumvents the problem of limited SNA focus and thereby presents a realistic scenario suitable for the communications infrastructure of the telecom companies (development of design technology for telecommunications service provider), is introduced.

II. Filter Methods

The UCS passes all data of the users to the filtering processes. These parameters are: communication partners and times, the communication media, resulting costs as well as possible profile information. If interaction occurs with a foreign actor from outside of the system limits of the UCS, which involves a previously unknown actor, this communication will be saved to the database as well. Doing so, all available data of the communication events is captured.

In the literature there are countless methods describing how to search available data bases according to problem oriented parameters. Most of these procedures have their origins in the optimization of business process in commercial fields. Others come from purely scientific queries, for example the researching of questions in sociology. In the UCS there exist the communication relationships of individual actors of the UCS amongst themselves and beyond the UCS. This data is spanned as a social network. In this way there is no differentiation made about the type of communication media that was used for a communications event that has taken place. In the social network there are clusters in which the partial totals of the actors are significantly more densely bound together compared to communications participants from other areas of the social network. These clusters are detected using a *k-Means-procedure* and the actors are appropriately assigned to them [17]. This class of procedure is a simple and fast method of cluster identification and is widely used in SNA applications. The number of clusters C_{count} is predetermined because the k-Means-procedure is hard partitioned. As a distance function of the k-Means-procedure the social proximity N_{soc} between

TABLE I
EXAMPLE PARAMETERS OF DIFFERENT COMMUNICATION MEDIA TO FEATURE EXTRACTION

communication media	user identification	location parameter	starting time	approx. expenses
eMail	MAIL FROM	initial MTA	delivery start time	amount of data
PSTN	phone number	prefix number	call start time	estimated charges
Cellular radio	phone number	home network	call start time	estimated charges
VoIP	SIP indentity	IP-Subnet	call start time	estimated charges

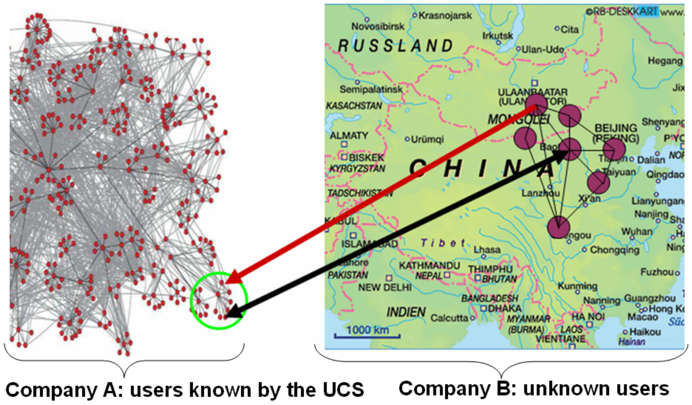

Figure 4. Example of a usage scenario with an incoming event

individual actors is used from (1).

At this point, the theory of operation will be explained with of a pithy example. In a fictive situation company A is collaborating with a company B during a product developement process. The two companies are far away from each other, one of them for example in China, see Fig. 4. The communication of the users in company A are protected with the new filters residing in their UCS. The heads of both development departments have a lively exchange over different communication media with each other, these events are indicated by the two black lines. Then an external developer hired by company B gets a problem definition to solve, for which he has to communicate with a developer of company A. So, both developers had not been in contact before and with the fact of the external hiring, traditional filters like (personal) listing procedures can not use common criterias of company B, e.g. domains in e-mail adresses or transmitted phone numbers. At the moment of receiving a communication event in company A's UCS, sent by the external developer, the filters know about the social proximity between company A's head of developement and company A's developer (actuall receiving). By the former communication behavior between the two developement heads in A and B, the filter system is aware of the heads A parameters in times of usage, distance of communication partners and costs of communication. When deciding about this communication event solely by tradional spam filters, the result could be unsure. And when deciding with former personal behavior of the developer in company A, the result could be unsure too. But now the corrective properties of company A's head of developement, which is detected to have a high social proximity to his developers, could be applied to the filtering process of this incomming communication event.

A. Data pre-processing

Each actor shows connections in the social network through the various communication processes with other actors from the range of available communications media. Out of the communication relationships of individual actors three characteristic relationship parameters are able to be derived by every communication process in every type of communications media available. Additionally, it is possible through the communication behaviour of closely connected actors to connect them to clusters and to in turn derive the three characteristic behaviour parameters of the clusters. These three behaviour parameters are the times h, at which communications procedures take place, the distance d between the participating actors and subsequently occurring costs c. The distance data are deduced from the technical parameters of the appropriate communications medium or communication system. Typically available communications media example parameters are shown in table I. Afterwards this makes procedures usable for position determination [20] [21]. The media specific costs due to a communication process can be deduced from telephone charges or the volume of transmitted data. In order to get the abstract comparable costs of a communication process from various communications media the cost parameters for these considerations are standardized units and therefore comparable in the sense used here.

Each actor is given a 3 tuple as parameter data for the three characteristic behaviour parameter, see (2). The elements of this make up the parameter and can be expressed through their index.

$$T_{m_i} = \begin{pmatrix} \vec{h}(m_i) \\ d(m_i) \\ c(m_i) \end{pmatrix} \tag{2}$$

Over a long period of time, a human actor displays time focal points upon which multiple communications have taken place. Capturing the time occurrences of the communication takes place in intervals in order to capture phases of increased communications incidences. This method is presented as a bar chart (see fig 5). A class wide of 60 minutes is proposed. Thereby a compromise between cancellation and complexity is given. Consequently a relationship is given between the division of the communications events in the different intervals and information about preferred communications points in time. As indicated in (3), the communications events are assigned to the 24 elements of the time vector $\vec{h}$.

$$\vec{h}(m_i) = \begin{pmatrix} \sum_{k=0}^{1} E_k(m) \\ \vdots \\ \sum_{k=23}^{24} E_k(m) \end{pmatrix} \tag{3}$$

The captured position of the actors m_i, at every communications event, and his corresponding partner actor is processed to the distance d. In this way a prototypical range of all communications can be calculated out of the number of all communication processes. This takes place through the mathematical mean, here the distance of all communication processes totalled is divided by the set, see (4).

$$d(m_i) = \frac{\sum_{k=1}^{E_m(Count)} E_{mk}(Distance)}{E_m(Count)} \tag{4}$$

Every communications instance has a cost c applied to it, in order to be able to classify the value of a communications instance, see

(5). Typically, this value is generated from the sending actor or his approximate surroundings. From the sum of all communication instances a prototypical cost value of the communication is calculated. This is represented through the mathematical mean, here the costs of all communication instances are totalled and divided by the set (see equation 4).

$$c(m_i) = \frac{\sum_{k=1}^{E_m(Count)} E_{mk}(Costs)}{E_m(Count)} \tag{5}$$

Subsequently the actor specific first part of the pre-processing clusters are searched for in the social network. For this the k-Means-procedure with the distance function given in equation 1 is used. Thereby the actors are connected to clusters which appear, through the communication behaviour, to be closely linked. After the clustering is carried out, each cluster C_i is assigned with the actors $C_{i,<acteurs>}$. According to the actors parameters three cluster parameters $\vec{h}(C_i)$, $d(C_i)$ and $c(C_i)$ are represented. These result from the arithmetic mean of the actor's parameters in the cluster. Thereby the parameter values of the actors in the cluster are added and divided by the set of the actors in the appropriate cluster. In this way the prototypical actor value for the cluster is generated. This 3 tuple is shown in (6).

$$T_{C_i} = \begin{pmatrix} \vec{h}(C_i) \\ d(C_i) \\ c(C_i) \end{pmatrix} \tag{6}$$

The captured values T_{m_i} and T_{C_i} can have their lines addressed through the index.

B. Evaluating the communications incidences

With the gathered and pre-processed data from the observed user behaviour, a simple test for spam or ham can be carried out upon the arrival of an external communication event E_r, on an actor in the UCS m_i. Thus, the actors and the cluster specific parameters are viewed as equal valued statements. With the linking up of individual behaviours, the affiliated actors and through the cluster analysis uncovered connection, the data basis for the decision is considerably enlarged and possible evaluation errors can be minimized. The cluster bound actors are socially affiliated and display similar communications behaviour. In this way the cluster specific statement confirms or corrects the actor specific statement. The ratio, between the new and individual communication event $current$ and the previously determined average value all, will be calculated for each of the three parameters $\vec{h}$, d and c. To evaluate a single communication event a non-naive approach is chosen, which means the result of the relationship comparison drives against the value *1.0.* This means the smaller the result's absolut value of the subtraction, the lower the probability of spam. This is represented by a subtraction of the evidence of a relationship comparison of one. In (7), the number of communication events in the class (time slice) of the current communication process $h_{current}$ is compared against all other classes according to the number of E_m contained.

$$h_{result}(E_r(m_i)) = 1.0 - \frac{\frac{T_{m_i}(1,current)}{T_{m_i}(1,all)} + \frac{T_{C_i}(1,current)}{T_{C_i}(1,all)}}{2} \tag{7}$$

In (8) the mean distance value from the recipient actor and his cluster are tested against the value of the received communication events.

$$d_{result}(E_r(m_i)) = 1.0 - \frac{\frac{T_{m_i}(2,current)}{T_{m_i}(2,all)} + \frac{T_{C_i}(2,current)}{T_{C_i}(2,all)}}{2} \tag{8}$$

In (9) the mean cost value of the previous communication of the recipient actors and their clusters are tested against the value of received communication processes.

$$c_{result}(E_r(m_i)) = 1.0 - \frac{\frac{T_{m_i}(3,current)}{T_{m_i}(3,all)} + \frac{T_{C_i}(3,current)}{T_{C_i}(3,all)}}{2} \tag{9}$$

C. Summary of the filter results

In order to summarize the three different filter results into a result value, a specific term must be used that allows it on the one hand to contain the total result, and on the other hand takes into account the characteristics of the individual filter levels. According to size, or rather construction of the social network in a database, various individual results can be achieved using evaluations metrics. A simple scoring method is based on experience values existing above the reliability of individual metrics in the usage context of the UCS. If it is recognized that the participant results invalidate, the spam level of a communication instance $E_{i,r}$, the weighting factors a, b and c not equal to 1.0 can be chosen. The individual metric results are multiplied with the scale factor and the product totalled.

$$Spamlevel(E_{i,r}) = a * h_{result} + b * d_{result} + c * c_{result} \tag{10}$$

For making a decision about a current $E_{i,r}$, whether it is spam or not, a threshold must be defined. The value to be used here is an individual nature, according to whichever risk of *false positives* appears to be acceptable.

D. Filter Position within a total context

The results from the filters described here can not reliably decide on a positive spam detection alone. Because of the multiplicity of possible connections a result value is only an additional indicator in the collection of all evaluation processes. Therefore a linking with the other (media specific) filters of individual communications media is allways necessary. Typically there are the hard criteria, such as firewall or listing procedures, which without cooperation with other filters reach a valid conclusion and soft criteria, such as content evaluation processes. For the final evaluation result of a communications process the results from all filter processes in this paper must be run together with the other processes. Typically a weighting according to reliability as well as personal settings of the associated communication participant will take place in the UCS.

III. Further Aspects

A telephone system's calculations data is taken and investigated for the characteristics of the three parameters $\vec{h}$, d and c. The data contains only the call placed by internal participants to unknown external numbers. The capture resulted originally for the purposes of billing only, not for the new procedure presented here. According to German data privacy law, the collection of unnecessary personal data is not allowed without agreement of every affected participant. There are no telephone calls between two or more known internal participants available, so that no social network in the appropriate sense can be spanned. The pure simulation of an artificial user group, in which each user generates the three parameters ($\vec{h}$, d and c) by random processes, would lead to results that are far away from reality. The following results presented in Fig. 4 to 6 originate from October 2009 data, the resulting analysis was carried out with standard tool boxes from Mathworks MatLab. There are 9298 communication events from 408 different call numbers available, a total of 42,211 charge units were used with an average speaking length of 287.48 seconds. The Fig. 5 to 7 were created using all available data thereby representing the total relationships of all participants. Clearly recognizable are working hours and weekends, as during that time little appreciable private call activity at the university takes place, see Fig. 5.

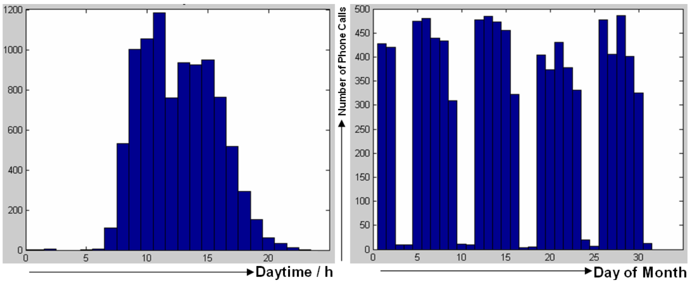

Figure 5. Call distribution over time of day and month

A large proportion of calls are shorter than five minutes and involve minimal cost, based on the charging units, see Fig. 6.

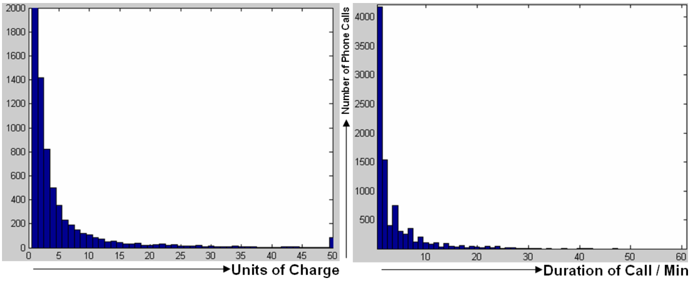

Figure 6. Call distribution over fee and time

The international calls are distributed over only a few target countries. This characteristic could be derived to a criterion for filtering by distance values, see Fig. 7.

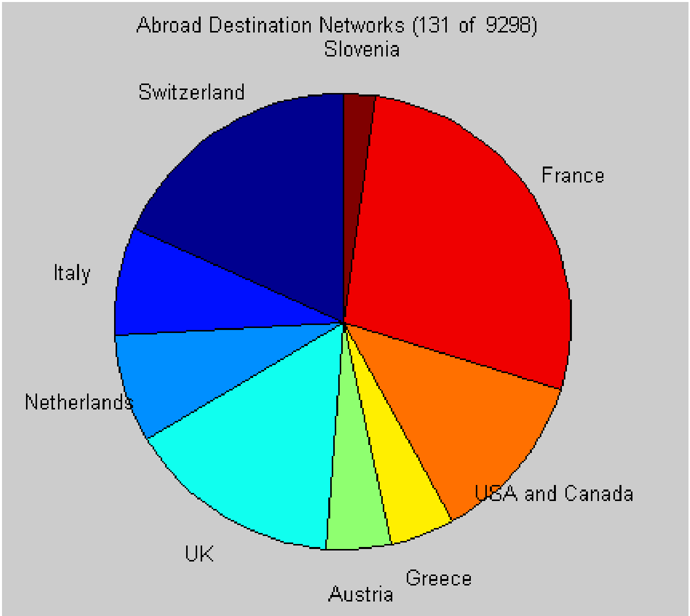

Figure 7. Abroad destination networks over a month

Through the broad range of the appropriately investigated parameters it is evident, that the procedure introduced here can give strongly conclusive results.
In order that these methods do not have to only be used with external databases an experimental UCS is being worked on at Pforzheim University, which unites the communications media of telephony and e-Mail and offers the possibility of the analysis introduced here. It will therefore be possible to evaluate the described procedures in a situation approximating reality and to variably test the parameters of different evaluation processes.

IV. Final remark

Proof of the effectiveness of the presented methods can only, untill now, be given from theoretical experiments as at the moment no data fulfilling the assumed conditions (social network capable) for this theory, is available. Because the decision procedures are based using the three parameters in (7), (8) and (9) on hard limitations, incorrect decisions on these limitation values are not unlikely. Real enviroments impose to implement an imprecise decision threshold, for example through the simple one-dimensional variance $VAR(X) = \sum(X - E)^2$ of the appropriate parameter X, with the arithmetical mean as expectation value E. The effectiveness of constant purported and thereby potentially suboptimal number of clusters through purported quantities of hard partitioned k-Means-procedure is also to be investigated using the available applicable data. Both input values from position and cost estimates will show different exactitude according to origin and communication instance. In combination with several relationship parameters and their histories these parameter input characteristics should be insignificant.
The procedure presented in this paper from three captured parameters of all conceivable communication media (time use behaviour, distance from communication partner and incurred - if only abstract - costs) in combination with individual and sum total behaviour as a reciprocal correction is a *new type* of idea in the battle against spam.

Acknowledgment

This research was made possible through Professor Dr.-Ing. Frank Niemann (Engineering) and the Pforzheim University as well as the Federal Ministry for Education and Research. The data represented in this paper does not necessarily represent the opinion of the Pforzheim University or the Federal Ministry for Education and Research.

References

[1] The Spamhaus Project, *The Definition of Spam*, February 2010, Checked September 2010 http://www.spamhaus.org/definition.html.
[2] Cisco IronPort, *2008 Internet Security Trends - A report on Emerging Attack Platforms for Spam, Viruses and Malware*, Checked September 2010 http://www.ironport.com/securitytrends/.
[3] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and Stefan Savage, Proceedings of the 15th ACM CCS, *Spamalytics: An Empirical Analysis of Spam Marketing Conversion*, October 2008, Checked September 2010 http://www.icsi.berkeley.edu/pubs/networking/2008-ccs-spamalytics.pdf.
[4] J. Klensin, *RFC 5321: Simple Mail Transfer Protocol (SMTP)*, October 2008, Checked September 2010 http://tools.ietf.org/html/rfc5321.
[5] Wikipedia, *Greylisting*, 4th February 2010, Checked September 2010 http://en.wikipedia.org/w/index.php?title=Greylisting&oldid=341929685.
[6] Rhyolite Software LLC, *Distributed Checksum Clearinghouses*, Summer 2008, Checked September 2010 http://www.dcc-servers.net/dcc/.
[7] Dr. S. Ritterbusch, *Die Mathematik des Bayes Spamfilters*, Checked September 2010 http://www.math.kit.edu/iag1/~ritterbusch/seite/spam/de.
[8] Cisco Systems Inc., *Description of the SenderBase Network*, Checked September 2010 http://www.senderbase.org/about.
[9] Cisco IronPort, *The SenderBase Network - Overview*, Checked September 2010 http://www.ironport.com/pdf/ironport_senderbase_overview.pdf.
[10] Federal Trade Commission, *The CAN-SPAM Act: A Compliance Guide for Business (Facts for Business)*, September 2009, Checked September 2010 http://www.ftc.gov/bcp/edu/pubs/business/ecommerce/bus61.shtm.

[11] The German Federal Ministry of Justice, *Gesetz gegen den unlauteren Wettbewerb*, 2004, Checked September 2010 http://www.gesetze-im-internet.de/englisch_uwg/index.html.

[12] The Spamhaus Project, *TRegister of Known Spam Operations (ROKSO)*, February 2010, Checked September 2010 www.spamhaus.org/rokso/.

[13] J. Rosenberg, et. al., *RFC 3261: Session Initiation Protocol (SIP)*, June 2002, Checked September 2010 http://tools.ietf.org/html/rfc3261.

[14] Heise-Newsticker, *Erste größere Attacke gegen deutsche VoIP-Nutzer*, September 2008, Checked September 2010 http://www.heise.de/security/meldung/Erste-groessere-Attacke-gegen-deutsche-VoIP-Nutzer-207400.html.

[15] D. Jansen, *Einfuehrung in die Netzwerkanalyse: Grundlagen, Methoden, Forschungsbeispiele*, August 2006 (3. Auflage), Vs Verlag,
ISBN-13: 978-3531150543.

[16] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining (Adaptive Computation and Machine Learning)*, October 2001, The MIT Press ISBN.

[17] H.-Y. Lam and D.-Y. Yeung, *A Learning Approach to Spam Detection based on Social Networks*, 2007, CEAS07 Fourth Conference on Email and AntiSpam, September 2010 http://www.ceas.cc/2007/papers/paper-81.pdf.

[18] P. O. Boykin and V. Roychowdhury, *Personal Email Networks: An Effective Anti-Spam Tool*, April 2005, IEEE Computer, Vol. 38, No. 4, pages 61-68.

[19] The Spamhaus Project, *The World's Worst Spam Producing Countries*, March 2010, Checked September 2010 http://www.spamhaus.org/statistics/countries.lasso.

[20] Wikipedia Encyclopedia, *Signaling System 7*, April 2010, Checked September 2010 http://en.wikipedia.org/w/index.php?title=Signaling_System_7&oldid=346781094.

[21] J. A. Muir and P. C. van Oorschot (Carleton University, Technical Report), *Internet Geolocation and Evasion*, April 2006, Checked September 2010 http://www.ccsl.carleton.ca/~jamuir/papers/TR-06-05.pdf.