Unwanted Traffic Control via Global Trust Management

Zheng Yan
Department of Communications and Networking, Aalto University
Espoo, Finland
XiDian University, Xi’an, China
zheng.yan@aalto.fi

Raimo Kantola
Department of Communications and Networking
Aalto University
Espoo, Finland
raimo.kantola@tkk.fi

Yue Shen
Department of Communications and Networking
Aalto University
Espoo, Finland
yue.shen@aalto.fi

***Abstract*— People’s life has been totally changed by the fast growth of Internet. It provides an incentive platform for many killer services and applications. However, it also offers an easy channel to distribute various contents that could be unwanted by users. This paper proposes a generic unwanted traffic control solution through global trust management. It can control unwanted traffic from its source to destinations according to trust evaluation. Simulation based evaluation shows that the solution is effective with regard to accuracy, efficiency and robustness against a number of malicious attacks.**

Keywords- unwanted traffic filtering; trust; trust managemen; reputation

I. INTRODUCTION

Internet, as a killer application, has become the backbone of remote communications, networking, computing and services. It carries a vast range of information resources and services, such as the World Wide Web and email. It also gives a birth to a wide range of applications, e.g., Voice over Internet Protocol (VoIP), Internet Protocol television (IPTV), Instant Messaging (IM), E-Commerce, Blogging, and social networking. However, at the same time as people collect information from the Internet, they could also unwanted traffic, such as malware, spam, spyware, intrusion attempts, and unsolicited commercial advertisement or contents, etc. Those unexpected or harmful information could intrude people’s devices, occupy device memory spaces, waste their time and irritate their usage experience. Some malicious traffic (e.g., a virus) has a fast infection speed and thus can be spread over the network quickly, but costs little on its source.

The unwanted traffics burden both users and internet service providers, however their source could benefit from propagating commercial advertisements for its business customers. In this paper, we propose to control unwanted traffic from its receivers to its source based on trust evaluation on each system entity and past unwanted traffic detection behaviors. We try to build an incentive structure and a global trust management system that includes all Internet Service Providers (ISPs), their subscribers (i.e., hosts) and a newly introduced global trust operator (GTO) to evaluate each system entity’s trust in order to decide how to control unwanted traffic. The trust contains two parts: one is the global trust that indicates if the entity is the source of unwanted traffic; the other is the detection trust that specifies the detection performance of each entity.

Concretely, we evaluate each involved system entity’s global trust in order to figure out if the traffic from it should be controlled for a receiver. The system entity can be a host or an ISP. The evaluation is based on both unwanted traffic detection reports from the hosts and traffic monitoring and check at ISP. Unlike prior arts [17], our solution reduces the overhead of ISP traffic monitoring by triggering this event according to the analysis of detection reports. In order to overcome potential attacks, we apply both global trust and detection trust to certify the reports from the host and ISP. Our design aims to provide a generic solution for different unwanted traffics over Internet, which is efficient to control unwanted traffic and robust to overcome a number of system attacks. The system is developed for the purpose of changing the current unwanted traffic ecosystem by turning the cost of unwanted traffic to its source based on the evaluation of trust. It provides evidence to decide the charge for unwanted traffic. Also, the sender side ISP can use the trust evaluation result to take an admistrative action against the offending sender. Although there have been a number of trust and reputation mechanisms proposed for controlling spam, spim (i.e., Instant Messaging spam), SPIT (Spam over Internet Telephony) and web pages [6, 8, 9, 10, 11, 14, 15, 17, 18, 21, 22, 23, 25], to our knowledge, such a generic solution as we propose in this paper is still lacking in the literature.

The rest of the paper is organized as follows. Section 2 gives a brief overview of related work. Section 3 introduces a global trust system structure followed by a procedure to globally control unwanted traffic. The algorithms used in the system are described in Section 4. We further evaluate the effectiveness of the designed algorithms through simulations in Section 5. Finally, conclusions and future work are presented in the last section.

II. BACKGROUND AND RELATED WORK

A. *Unwanted Traffic Detection Technology*

There are quite a number of existing anti-spam techniques and applications, e.g., whitelists/blacklists; header/content checks and rule-based filtering (e.g., SpamAssassin [26]); Bayesian analysis (e.g., SpamBayes [27]); sender authentication (e.g., Sender Policy Framework [28], Yahoo DomainKeys [29], etc.); challenge/response (e.g., TMDA [30]), Blackhole listing (e.g., SORBS, Kelkea MAPS [30]), and distributed checksums (e.g., DCC, Vipuls Razor [30]). A problem with whitelists and blacklists is that they leave a sizable set of senders in the middle of the spectrum that are not classified. In addition, we argue that spam filters, intrusion detection systems, and firewalls are reactive, i.e., defensive tools. They are good at collecting evidence on suspect behavior of senders. Although various tools exist, we still lack global control and management on unwanted traffic. In our opinion, the reports of various detection tools can play as a valuable input to the global control of unwanted traffic. As spammers started using fast changing botnets, randomizing and obfuscating content in their messages, the above technologies quickly became ineffective. An efficient method is expected. The solution proposed in this paper partially depends on evidence collected from each entity involved in the global Internet system. The collection could be based on various existing tools or a user's behaviors or ratings.

B. *Unwanted Traffic Control via Trust Management*

For unwanted traffic control, autonomic trust management aims to control or filter traffic automatically based on the trust relationship between the traffic source and its receiver [3, 31]. A number of solutions were proposed to control unwanted traffic via trust and reputation mechanisms. Most existing unwanted traffic control systems based on trust and reputation mechanisms target on email spam.

A distributed architecture and protocol for establishing and maintaining trust between mail servers was proposed in [14]. The architecture is a closed loop control system that can be used to adaptively improve spam filtering by automatically using trust information to tune the threshold of such filters. A layered trust management framework was proposed in order to help email receivers eliminate their unwitting trust and provide them with accountability support [11]. In [15], IPGroupRep clusters the senders into different groups based on their IP addresses and computes the reputation value of each group according to email receiver's feedback on the messages sent from them. The reputation value can be used to indicate whether an incoming message is spam or not. However, the above solutions did not consider malicious attacks on the proposed system, e.g., wrong/malicious feedbacks.

MailTrust filters out dishonest feedbacks to obtain an accurate trust value of each mail server [21]. The credibility-based reputation generation is similar to the detection trust in our solution. But MailTrust is a distributed reputation system, while ours is a centralized one, based on GTO. In [17], email senders' behavior was analyzed in order to figure out spammers. This method is a predictive approach based on static statistical analysis, which cannot be applied into an unwanted traffic control system at runtime, like ours. Thus it is not efficient to control fast spreading botnets. A multi-level reputation-based greylisting solution was proposed to improve the efficiency of traditional greylisting anti-spam methods by significantly reducing the transfer delay of messages caused by the additional greylisting level [18]. Comparing to the above work, the trust evaluation in our solution is not only based on each host's detection reports, but also the monitored behavior of unwanted traffic source at ISP.

Highly related to our work, a framework for a reporter-based reputation system for spam filtering was proposed to filter spam [6]. The system includes a trust-maintenance component, in which users gain and lose reputation, depending on their spam-reporting patterns (similar to the detection trust in our solution). The filtering component uses the reports of highly reputable reporters for spam removal. This work focused on large quantities of highly similar spam (i.e., a campaign) sent within a relatively short period of time. The authors did not discuss its applicability on other types of unwanted traffic. Further study and analysis are needed to control various types of unwanted traffics over Internet.

A number of algorithms attempted to overcome web page spam, such as PageRank [23]. Most link-based anti-spamming algorithms are based on observed features, in which spam pages are different from reputable ones [9]. Some anti-spamming algorithms utilize users' implicit or explicit feedback in assisting the page ranking, such as BrowseRank [24]. TrustRank and its variations firstly select a certain number of seeds for experts' manual evaluation and then propagate trust or distrust through links from the seed sets [10, 25]. Obviously, this mechanism is not suitable for controlling other types of unwanted traffic.

Voice spamming, SPIT was studied in [8]. Unlike spam in e-mail systems, VoIP spam calls have to be identified in real time. Thus, many techniques devised for e-mail spam detection are not practical for SPIT. To overcome this challenge of blocking a spam call before telephone rings, Kolan and Dantu proposed a multi-stage, adaptive spam filter based on presence (location, mood and time), trust, and reputation to detect SPIT. But, this solution is specific for SPIT, which is hard to be widely applied into other scenarios. However, this method can play as a specific detection tool for SPIT in our proposed system.

In the context of instant messaging, a trust and reputation based anti-SPIM method was proposed in [22]. This method integrated trust and reputation with black-list/white-list techniques. Since the method is proposed for IM spam (i.e., SPIM), trust and reputation generation is based on IM social networking, which cannot be directly applied into other application scenarios.

In summary, literature still lacks a generic unwanted traffic control mechanism, which should be efficient, accurate, robust and economic to be flexibly embedded into the current Internet architecture to control various types of unwanted traffic.

III. System Structure and Unwanted Traffic Control Procedure

A. *Assumptions and Requirements*

Our research holds a number of assumptions based on existing work as described below [32]:

1. Identity assumption: A source of unwanted traffic and its receiver in most cases can be identified with the accuracy

of an IP address prefix or a NAT (network address translation) outbound IP address when a NAT hides the source host or receiver host itself. Meanwhile, each content/traffic can be identified based on its hash code.

2. GTO assumption: A Global Trust Operator behaves as an authorized trustworthy party to collect trust evidence and conduct global trust evaluation on different system entities. We assume that a secure and dependable communication channel is applied in the system for unwanted traffic reporting and controlling.
3. Traffic assumption: We assume that the unwanted traffic is sourced from a host and targets other hosts via its local ISP. (In this case, web page spam could be an exception.)
4. Detection assumption: We assume that the unwanted traffic can be detected at the host either manually or automatically with some installed toolkits.
5. Forward assumption: We assume that each local ISP timely and honestly forwards the reports from its host to the GTO in a secure way.

We recognize the key desirable properties of a global trust management system for unwanted traffic control as below:

(1) Timely/efficient and accurate recognition of unwanted traffic;
(2) Automatic maintenance of trust for each system entity;
(3) Robustness against various attacks on the system.

B. System Structure

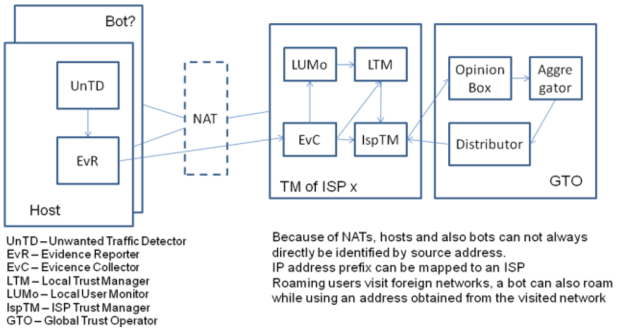
Bot?
UnTD
EvR
Host
NAT
LUMo
LTM
EvC
IspTM
TM of ISP x
Opinion Box
Aggre gator
Distributor
GTO
UnTD – Unwanted Traffic Detector
EvR – Evidence Reporter
EvC – Evicence Collector
LTM – Local Trust Manager
LUMo – Local User Monitor
IspTM – ISP Trust Manager
GTO – Global Trust Operator
Because of NATs, hosts and also bots can not always directly be identified by source address.
IP address prefix can be mapped to an ISP
Roaming users visit foreign networks, a bot can also roam while using an address obtained from the visited network

Figure 1. System structure

Fig. 1 shows the structure of the global trust management system for unwanted traffic control. At the host side, it embeds an Unwanted Traffic Detector (UnTD), which can be any unwanted traffic detection toolkits for different kinds of contents (e.g., Email, VoIP, IM, web pages and intrusions). An Evidence Reporter (EvR) at the host reports the unwanted traffic detection results to its local ISP, where an Evidence Collector (EvC) collects the reports. Each ISP in our system has a trust manager (TM). It contains a number of functional blocks in order to do unwanted traffic control. Concretely, a Local User Monitor (LUMo) is applied to monitor the traffic sourced from a local system entity. A Local Trust Manager (LTM) conducts trust analysis based on the evidence collected from local hosts and/or the input from LUMo. The analysis results are used to trigger traffic monitoring at ISP (LUMo) and traffic similarity check. An ISP Trust Manager (IspTM) is responsible for evaluating trust locally, transferring the results from the LTM to GTO, requesting GTO for global trust evaluation and unwanted traffic control, and receiving the global trust value of the requested entity and a blacklist of unwanted traffic sources, as well as personalized traffic control decision from GTO. At the GTO side, an Opinion Box is used to securely store trust evidence and information that are used to evaluate global trust of each entity and make an unwanted traffic control decision at an Aggregator. At GTO, a Distributor is applied to collect trust evidence and information, receive requests from ISPs and distribute the decisions of GTO to ISPs.

The application condition of the proposed system is that the unwanted traffic is sourced from a host in Internet and sent to other hosts through its local ISP. It is applicable to various unwanted traffics, such as spam, spim, sipt, and so on. Web page spam could be an exception. But if the web page or its link is sent in this way to other hosts, our system is still applicable.

C. A Global Unwanted Traffic Control Procedure

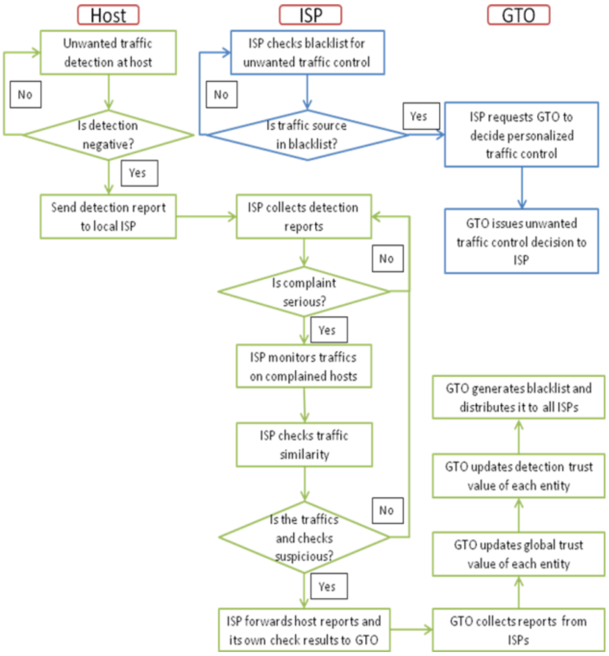
Host
ISP
GTO
Unwanted traffic detection at host
No
Is detection negative?
Yes
Send detection report to local ISP
ISP checks blacklist for unwanted traffic control
No
Is traffic source in blacklist?
Yes
ISP requests GTO to decide personalized traffic control
GTO issues unwanted traffic control decision to ISP
ISP collects detection reports
No
Is complaint serious?
Yes
ISP monitors traffics on complained hosts
ISP checks traffic similarity
No
Is the traffics and checks suspicious?
Yes
ISP forwards host reports and its own check results to GTO
GTO collects reports from ISPs
GTO updates global trust value of each entity
GTO updates detection trust value of each entity
GTO generates blacklist and distributes it to all ISPs

Figure 2. A global unwanted traffic control procedure

We propose a procedure to conduct unwanted traffic control via global trust management based on the above system structure, as shown in Fig. 2.

Concretely, the host device detects the unwanted traffic and reports to its local ISP if the detection is problematic. The ISP collects the complaint reports from hosts. If the complaint is serious, ISP does traffic monitoring on suspicious hosts and content similarity check. If the above checks are abnormal or suspicious, ISP forwards host reports and its own check results to GTO. The GTO collects the reports from ISPs all over the world. It aggregates all collected information and evaluates each system entity (host or ISP)'s global trust and thus detects the source of unwanted traffic and send blacklists to ISPs. Particularly, our proposed system can provide personalized unwanted traffic control. In this case, the ISP sends a request to the GTO if it finds traffic from a host in the blacklist is sent to some destination hosts. Based on the past detection reporting behaviors, the GTO can decide if the traffic should be controlled for each destination host.

D. Trust Evaluation

In the proposed system, each host uses existing tools or manual ways to detect unwanted traffic and reports to its local ISP. Each ISP runs Local User Monitoring and evaluates local trust in other entities connected directly to the ISP network. There might be several global trust operators that must cooperate. In the current design, we stick to a single global trust operator at least for the time being.

The GTO is responsible for evaluating global trust of each system entity, generating a blacklist accordingly and making unwanted traffic control decisions. For generating a blacklist, each incident/piece of evidence collected is assigned a credibility or confidence value, which is the detection trust of its provider.

IV. ALGORITHMS

TABLE I. NOTATIONS

Symbol	**Description**
$f(x)$	The Sigmoid function $f(x)=\frac{1}{1+e^{-x}}$; used to normalize a value into (0, 1);
U_k	The system entity, it can be either an ISP or a host;
$tr_k(t)$	The traffic of host U_k at time *t*;
$d_t\{g(t)\}$	$d_t\{g(t)\}=\frac{g(t)-g(t-\tau)}{\tau}$, $(\tau \to 0)$; $g(t)$ is a function of variable *t*;
e_i^k	The *i*th content received by host U_k ;
$v_k^i(t)$	The possibility of content e_i^k being unwanted traffic indicated by U_k at time *t*; $v_k^i(t) \in [0,1]$
ut_k^t	The global trust of U_k at time *t*;
$s_i^k(t)$	The unwanted traffic detection value at time *t* by U_k about e_i^k ;
thr	The threshold of the host to report to ISP;
$thr0$	The threshold to trigger traffic monitoring at local ISP;
$thr1$	The threshold of ISP to report to GTO
φ_{sp}^k	The unwanted traffic indicator contributed by the ISP traffic monitoring on U_k ;
sim_i^k	The similarity of contents correlated to e_i^k ;
sim^k	The content similarity factor of U_k by considering all similar contents sent by U_k ;
$\theta(I)$	The Rayleigh cumulative distribution function to model the impact of integer number I ;
$sp_k^n(t)$	The unwanted traffic detection value about host U_k provided by the *n*th ISP SP_n at time *t*;
$rt_{k'}^t$	The contribution of reports from the hosts to the evaluation of $U_{k'}$'s global trust at time *t*;
$mt_{k'}^t$	The contribution of reports from the ISPs to the evaluation of $U_{k'}$'s global trust at time *t*;
dt_k^t	The detection trust of entity U_k at time *t*;
δ	The parameter to control the adjustment of dt_k^t ;
γ	The warning flag to record the number of bad detections;
μ	The parameter to control bad detection punishment;
$thr2$	The threshold to put an entity into the blacklist at GTO;
$thr3$	The threshold to determine on-off or conflict behavior attack;

Based on the above system design, we propose a number of algorithms to implement unwanted traffic control via global trust management. For easy of reference, Table I summarizes the notations used in this section.

A. Unwanted Traffic Detection at Host and Aggregation

Assume that a number of unwanted traffic detection tools are applied or installed at the user's device (host), they detect unwanted traffic and report the detection result to the host's local ISP.

1) Unwanted Traffic Reporting

The host U_k reports a received content e_i^k as unwanted at time *t* as $v_k^i(t)$. It is automatically sent to its local ISP if $v_k^i(t) \geq thr$.

The credibility of $v_k^i(t)$ is U_k 's global trust value ut_k^t. Thus the detection value $s_i^k(t)$ at time *t* by U_k about content e_i^k is described as:

$$s_i^k(t) = v_k^i(t) * ut_k^t . \tag{1}$$

The reports can be aggregated at ISP in order to decide whether traffic monitoring and check at ISP is needed. The aggregation is based on the following formula:

$$s_i(t) = \frac{\sum_k v_k^i(t) * ut_k^t}{\sum_k ut_k^t} \tag{2}$$

B. Traffic Monitoring at ISP

The purpose to monitor a host U_k 's traffic at its local ISP is to find the senders of unwanted traffic with such credibility that either administrative action can be taken by the ISP or contractual penalties can be imposed by the ISP on the sender. . This traffic monitoring on a specific host or ISP is triggered by a condition ($s_i(t) \geq thr0$) in order to save the running cost of ISP for unwanted traffic control. Particularly, it can detect an infected host who becomes a source of unwanted traffic due to infection. U_k can be any entity (either an ISP subscriber or other ISPs) that links to the ISP, thus its traffic can be monitored by the ISP. The traffic deviation of U_k at time *t* can be described as $d_t\{tr_k(t)\}$, where $d_t\{g(t)\}=\frac{g(t)-g(t-\tau)}{\tau}$, $(\tau\to 0)$; $g(t)$ is a function of variable *t*. $d_t\{tr_k(t)\}$ indicates the traffic changes of U_k, the bigger the changes, the more probability U_k is infected as an unwanted traffic source. Thus, an unwanted traffic indicator contributed by the ISP traffic monitoring on U_k is

$$\varphi_{sp}^k(t)=\left|1-2f\{d_t\{tr_k(t)\}\}\right|. \qquad (3)$$

Meanwhile, we also check the similarity of content sent out from U_k. For a set of similar sized traffics $E_k=\{e_i^k\}$, $i=\{1,......,I\}$, we calculate their similarity as

$$sim_i^k=\frac{\theta(I)}{I-1}\sum_{i'\neq i}^{I}\left(1-\left|e_i^k-e_{i'}^k\right|\right), \qquad (4)$$

where $\left|e_i^k-e_{i'}^k\right|$ is the difference between e_i^k and $e_{i'}^k$. It can be calculated based on a semantic relevance measure, such as the cosine similarity and the one described in [4]. Note that, we also consider *I*'s influence (i.e., the number of similar contents sourced from a host) by applying the Rayleigh cumulative distribution function:

$$\theta(I)=\left\{1-\exp\left(\frac{-I^2}{2\sigma^2}\right)\right\}, \qquad (5)$$

where $\sigma>0$, is a parameter that inversely controls how fast the number of similar unwanted traffic impacts on sim_i^k, it increases as I increases. Parameter σ can be set from 0 to theoretically ∞, to capture the characteristics of different scenarios. We set $\sigma=100$ in our simulations.

U_k could be the source of multiple *M* unwanted traffics, thus we have

$$sim^k=\frac{1}{M}\sum_M\left\{\frac{\theta(I)}{I-1}\sum_{i'\neq i}^{I}\left(1-\left|e_i^k-e_{i'}^k\right|\right)\right\}. \qquad (6)$$

Thus, the unwanted traffic detection value about U_k provided by SP_n at time *t* is:

$$sp_k^n(t)=\varphi_{sp}^k(t)*sim^k. \qquad (7)$$

ISP reports its monitoring result $\varphi_{sp}^k*sim^k$ to the GTO if $\varphi_{sp}^k\geq thr1$ or $sp_k^n(t)\geq thr1$. Meanwhile, ISP will honestly forward any unwanted traffic reports from the hosts to the GTO by adding the source ID of the complained traffic. Algorithm 1 is applied to monitor unwanted traffic at ISP.

Algorithm 1: Unwanted Traffic Monitor at ISP

```
Input:
  - $tr_k(t)$, $E_k=\{e_i^k\}$ $i=\{1,......,I\}$;
  - $U_k$ ($k=1,\ldots,K$).
For each complained $U_k$, do
  Monitor $U_k$'s traffic, calculate $\varphi_{sp}^k(t)=\left|1-2f\{d_t\{tr_k(t)\}\}\right|$;
  Calculate $sim^k$ and $sp_k^n(t)$;
  if $\varphi_{sp}^k\geq thr1$ or $sp_k^n(t)\geq thr1$
    Report it to Global Trust Operator (GTO).
Output: $sp_k^n(t)$, $n=(1,...,N)$.
```

C. *Unwanted Traffic Control at GTO*

The GTO evaluates each entity's trust based on collected reports from the hosts and ISPs in order to find the source of unwanted traffic. For the reports from the hosts, the GTO checks the source of the traffic thus find the identity of the complained host $U_{k'}$.

Obviously, U_k ($k=1,\ldots,K1$) could report e_i^k as an unwanted traffic for many times at different time t: $\{V_k^i\}=\{V_k^i(t)\}$. Considering the time's influence and potential ballot stuffing and on-off attacks, we pay more attention to the host's recent reports by introducing time decaying. For each system entity $U_{k'}$ (identified by its IP address) who is complained, we aggregate the reports from *K1* hosts who complained this source as below:

$$rt_{k'}^{t_p}=\frac{\sum_{k=1}^{K1}ut_k^{t_p}*v_k^i(t)*e^{-\frac{\left|t-t_p\right|^2}{\tau}}}{\sum_{k=1}^{K1}ut_k^{t_p}e^{-\frac{\left|t-t_p\right|^2}{\tau}}}, \qquad (8)$$

where t_p is the global trust evaluation time, τ is a parameter to control the time decaying, ($\tau=2$ in our simulations). We further consider the reports from *N* ISPs' monitoring and checks as the contributions of ISPs on $U_{k'}$'s global trust aggregation.

$$mt_{k'}^{t_p}=\frac{\sum_{n=1}^{N}ut_n^{t_p}*sp_{k'}^n(t_p)}{\sum_{n=1}^{N}ut_n^{t_p}} \qquad (9)$$

Thus, we evaluate the global trust value of the complained entity *k'* as:

$$ut_{k'}^{t_p}=ut_{k'}^{t_p}-rt_{k'}^{t_p}-mt_{k'}^{t_p}. \qquad (10)$$

Further considering the number of complainers, we have

$$ut_{k'}^{t_p}=ut_{k'}^{t_p}-\theta(K1)*rt_{k'}^{t_p}-\theta(N)*mt_{k'}^{t_p}. \qquad (11)$$

Algorithm 2 is used to conduct global trust evaluation and unwanted traffic control at ISP or GTO.

Algorithm 2: Global Trust Evaluation at GTO and

Unwanted Traffic Control

```
1. Input:
2.   - $sp_k^n(t_p)$, $v_k^i(t)$, ( $i = 1,......, I_{U_k}$ ).
3. For each complained system entity $k'$, do
4.    Calculate $rt_{k'}^{t_p}$, $mt_{k'}^{t_p}$, and $ut_{k'}^{t_p}$ based on (1)-(11), (1'), (7') - (9');
5.    If $ut_{k'}^{t} \leq thr2$, put $U_{k'}$ into blacklist
6. Output: blacklist $U_{k'}$.
7. Action: control traffic sourced from blacklist
```

D. Detection Trust: The Credibility of Detection

We need to evaluate the credibility of detection in order to fight against various attacks and malicious behaviors of hosts, for example, the host is attacked; the host is malicious; the detection tools installed in the host are broken or hacked; there is no trusted computing platform support at the host device; the detection tools are poor and the detection is not qualified. We further introduce detection trust to indicate the credibility of unwanted traffic detection, which is another dimension of trust.

The detection trust of entity U_k (either a host or an ISP) is generated at GTO as below:

If the detection reported by U_k doesn't match the final evaluation result, $y = -1$, and $\gamma{+}{+}$. Otherwise, $y = 1$. The detection trust dt_k^t of U_k at time *t* is:

$$dt_k^t = \begin{cases} dt_k^t + \delta y & (\gamma < thr3) \\ dt_k^t + \delta y - \mu\gamma & (\gamma \geq thr3) \end{cases} = \begin{cases} 1 & (dt_k^t > 1) \\ 0 & (dt_k^t < 0) \end{cases} \qquad (13)$$

Where $\delta > 0$ is a parameter to control the adjustment of dt_k^t. We further introduce a warning flag γ to record the number of bad detections. γ's initial value is 0. It is increased by 1 each time when a bad detection happens. $thr3$ is a threshold to indicate the number of malicious behavior attacks. $\mu > 0$ is a parameter to control bad detection punishment. In our simulation, we set $\delta = 0.05$, $\mu = 0.1$, and *thr3=5*. We set the initial value of dt_k^t as 0.5.

The detection trust can be adopted if no Trusted Computing Platform technology is applied in the devices of hosts and ISPs. Considering the detection trust, we have

$$s_i(t) = \frac{\sum_k v_k^i(t) * ut_k^t * dt_k^t}{\sum_k ut_k^t * dt_k^t} \qquad (1')$$

$$sp_k^n(t) = \varphi_{sp}^k(t) * sim^k * dt_n^t \qquad (7')$$

$$rt_{k'}^{t_p} = \frac{\sum_{k=1}^{K1} dt_k^{t_p} * ut_k^{t_p} * v_k^i(t) * e^{-\frac{|t-t_p|^2}{\tau}}}{\sum_{k=1}^{K1} dt_k^{t_p} * ut_k^{t_p} e^{-\frac{|t-t_p|^2}{\tau}}} \qquad (8')$$

$$mt_{k'}^{t_p} = \frac{\sum_{n=1}^{N} dt_n^{t_p} * ut_n^{t_p} * sp_{k'}^n(t_p)}{\sum_{n=1}^{N} dt_n^{t_p} * ut_n^{t_p}} \qquad (9')$$

V. Evaluation and Analysis

A. Simulation Settings and Evaluation Measure

We design a number of simulations to evaluate the feasibility and effectiveness of our proposed system with regard to accuracy, efficiency and robustness. In our simulations, we have a total of K=1000 hosts, N=5 ISPs. Each ISP has 200 hosts connected. There are L (=3, 5, 10, or 50) sources of unwanted traffic. All of them randomly select a number of hosts to send unwanted traffic within a time period. For a good host which is not attacked, it reports unwanted traffic in a good way. For a malicious host, it reports the unwanted traffic with a malicious pattern, e.g., (a) hide evidence attack - don't report the unwanted traffic; b) bad mouthing attack - report the traffic from a good source as unwanted; c) on-off attack - alternatively report normally or badly in order to hide its malicious behaviors. In our simulations, we assume that the unwanted traffics from the same source are identical. The initial global trust value of each system entity is 1; the initial detection trust value of each entity is 0.5. Table II provides the simulation settings of other system parameters.

TABLE II. SIMULATION SETTINGS OF SYSTEM PARAMENTERS

Symbol	Settings	Symbol	Settings
thr	0.8	σ	100
$thr0$	0.7;	τ	2
$thr1$	0.8;	δ	0.05
$thr2$	0.0001	μ	0.1
$thr3$	5;		

We adopt commonly used metrics in information retrieval, Recall (*R*), Precision (*P*) and F measure (*F*) to describe the performance of unwanted traffic control. We denote the number of entities that belong to the source of unwanted traffics (*SUT*) and are indeed detected as *SUT* as *x*; the number of entity that don't belong to *SUT* but are added to *SUT*, denoted as *y*; the number of entities that belong to *SUT* but are not detected as *SUT*, denoted as *z*. With these data we do a precision-recall evaluation. We define:

$$R = \frac{x}{x+z} \qquad (14)$$

$$P = \frac{x}{x+y} \qquad (15)$$

$$F = \frac{2PR}{P+R} \tag{16}$$

Obviously, $R, P, F \in [0,1]$. High recall, precision and F measure are desirable for good performance of the system.

B. Experiment 1: Accuracy of Unwanted Traffic Source Detection

We design Experiment 1 to test the accuracy of our system. In this experiment, we test the F measure in the following simulation condition: *there is no botnet, only (3, 5, 10, or 50) original independent unwanted traffic sources and all system entities are good.* Fig. 3 shows the experiment result. We observe that the system can accurately detect the unwanted traffic sources in an efficient way when they occupy no more than 5% of the system hosts (which is the normal case in practice).

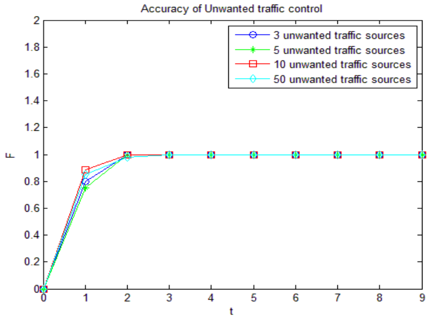

Figure 3. Accuracy of unwanted traffic source detection

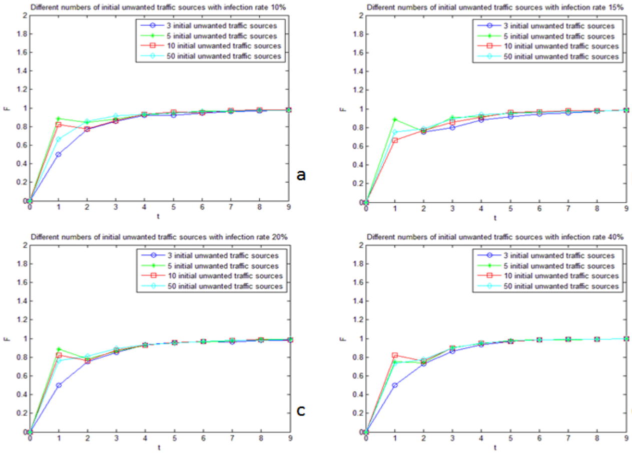

Figure 4. Efficiency of unwanted traffic source detection with different infection rates: (a) 10%; (b) 15%; (c) 20%; (d) 40%.

C. Experiment 2: Efficiency of Unwanted Traffic Source Detection

We design Experiment 2 to test the efficiency of our system. Efficiency can be reflected by detection speed/performance (number of time period), i.e., how fast the system can detect the sources of unwanted traffic. In this experiment, we test the F measure in the following simulation settings: *every time 10% (15%, 20%, and 40%) unwanted traffic destination hosts are infected and fall into a botnet.* Fig. 4 shows the experiment result. We observe that the system can efficiently detect the unwanted traffic sources (F reaches 1 within 9 time periods) even when 40% destination hosts are infected in the situation that the unwanted traffic sources occupy no more than 5% of the system hosts.

D. Experiment 3: Robustness of Unwanted Traffic Control

Malicious or hacked hosts could intentionally hack our system. We design Experiment 3 to test the robustness of our system. In this experiment, we test the F measure in the following simulation settings: *there is no botnet, only a number of original independent unwanted traffic sources with attacks raised by some malicious hosts*. We test two malicious attacks:(a) Hide evidence attack: malicious hosts hide detection evidence. The proportion of malicious hosts is 10%, 15%, 20%, and 40% of the whole hosts, respectively. The simulation result is shown in Fig. 5. We observe that our system performs very well against the hide evidence attack. The F measure can generally reach 1 within 4 time periods in the situation that the unwanted traffic sources occupy no more than 5% of the system hosts.

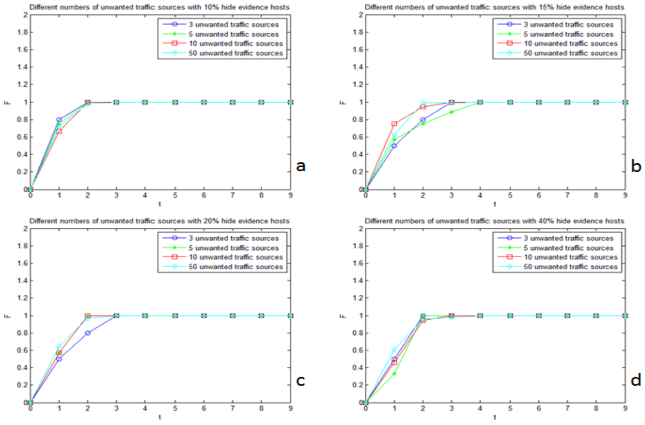

Figure 5. Robustness of unwanted traffic source detection with hide evidence attack: (a) 10% hide evidence hosts; (b) 15% hide evidence hosts; (c) 20% hide evidence hosts; (d) 40% hide evidence hosts.

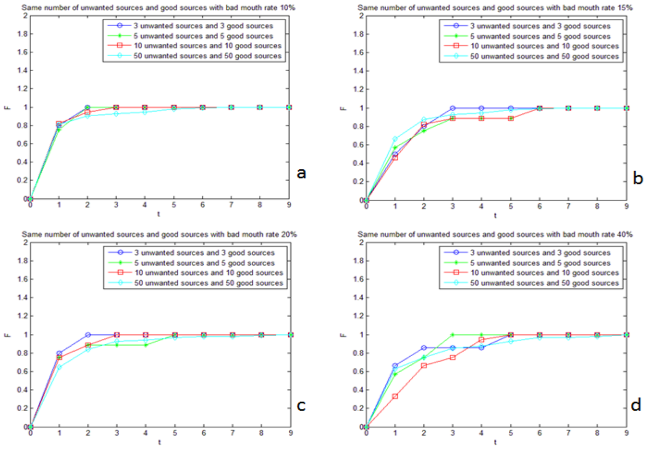

Figure 6. Robustness of unwanted traffic source detection with bad mouthing attack: (a) 10% bad mouth hosts; (b) 15% bad mouth hosts; (c) 20% bad mouth hosts; (d) 40% bad mouth hosts.

(b) Bad mouthing attack: malicious hosts intentionally frame a good traffic as unwanted. We simulate some good traffic in the system and test the situations that the proportion of bad mouth hosts is 10%, 15%, 20%, and 40% of the whole hosts, respectively. The simulation result is shown in Fig. 6. We observe that our system performs well against the bad mouthing attack. The F measure can generally reach 1 within 9 time periods in the situation that the unwanted traffic sources occupy no more than 5% of the system hosts.

VI. Conclusions and Future Work

The literature still lacks a generic and effective unwanted traffic control solution over Internet. This paper proposed a generic unwanted traffic control system based on global trust management by introducing a global trust operator. We designed a number of algorithms that can be adopted by the system to control unwanted traffic. The simulation results show our system's effectiveness with regard to accuracy, efficiency and robustness against a number of malicious attacks. Our paper contributes to the literature in two folds: (1) it proposed a global and generic unwanted traffic control system over Internet; (2) it automatically maintains each system entity's global trust and detection trust in a dynamic way, thus it can overcome a number of malicious attacks.

Regarding the future work, we will further improve the system by investigating its performance with practical data and solving practical issues. We will attempt to apply it into practice by studying its economy impact on current Internet ecosystem and exploring a proper business model that can be accepted by system stakeholders.

References

[1] Z. Yan, Trust Management for Mobile Computing Platforms, PhD dissertation, Dept. of Electrical and Communication Eng., Helsinki Univ. of Technology, 2007.

[2] Y. Sun, W. Yu, Z. Han, and K. J. R. Liu, "Information theoretic tramework of trust modeling and evaluation for ad hoc nNetworks," IEEE Journal on Selected Area in Communications, vol. 24, no.2, pp. 305-317, 2006.

[3] Z. Yan and C. Prehofer, "Autonomic trust management for a component based software system," IEEE Transactions on Dependable and Secure Computing, vol.8, no.6, pp. 810-823, 2011. doi: 10.1109/TDSC.2010.47

[4] J. Wang, F. Wang, Z. Yan, and B. Huang, "Message receiver determination in multiple simultaneous IM conversations", IEEE Intelligent Systems, vol.26, no.3, pp. 24-31, 2011.

[5] Z. Yan (Ed.), Trust Modeling and Management in Digital Environments: from Social Concept to System Development, IGI Global, 2010.

[6] E. Zheleva, A. Kolcz, and L. Getoor, "Trusting spam reporters: a reporter-based reputation system for email filtering", ACM Transactions on Information Systems, vol. 27, no. 1, Article 3(27), December 2008.

[7] L. Nie, B. Wu, and B. D. Davison, "Winnowing wheat from the chaff: propagating trust to sift spam from the web", SIGIR '07, pp. 869-870, July 2007.

[8] P. Kolan and R. Dantu, "Socio-technical defense against voice spamming", ACM Transactions on Autonomous and Adaptive Systems, vol. 2, no. 1, Article 2(44), March 2007.

[9] X. Zhang, B. Han, and W. Liang, "Automatic seed set expansion for trust propagation based auti-spamming algorithms", WIDM'09, pp. 31-38, November 2009.

[10] B. Wu, V. Goel, and B. D. Davison, "Topical TrustRank: using topicality to combat web spam", WWW '06, pp. 63-72, May 2006.

[11] W. Liu, S. Aggarwal, and Z. Duan, "Incorporating accountability into internet email", SAC '09, pp. 975-882, March 2009.

[12] P. Kumaraguru, A. Acquisti, and L. F. Cranor, "Trust modeling for online transactions: a phishing scenario", PST'06, pp. 1-9, October 2006.

[13] P. Varalakshmi, S. Thamarai Selvi, S. Monica, and G. Akilesh, "Securing trustworthy three-tier grid architecture with spam filtering", First International Conference on Emerging Trends in Engineering and Technology, pp. 396-399, 2008.

[14] J. McGibney and D. Botvich, "A trust overlay architecture and protocol for enhanced protection against spam", The Second International Conference on Availability, Reliability and Security, ARES 2007, pp. 749-756, 2007.

[15] H. Zhang, H. Duan, W. Liu, and J. Wu, "IPGroupRep: A novel reputation based system for anti-spam", Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, pp. 513-518, 2009.

[16] C. D. Curran, "Combating spam, spyware, and other desktop intrusions: legal considerations in operating trusted intermediary technologies", IEEE Security & Privacy, vol. 4, no. 3, pp. 45-51, 2006.

[17] Y. Tang, S. Krasser, Y. He, W. Yang, and D. Alperovitch, "Support vector machines and random forests modeling for spam senders behavior analysis", IEEE GLOBECOM, pp. 1-5, 2008.

[18] A. G. K. Janecek, W. N. Gansterer, and K. A. Kumar, "Multi-level reputation-based greylisting", ARES08, pp. 10-17, 2008.

[19] D. Polz, and W. N. Gansterer, "Trustnet architecture for e-mail communication", DEXA '09, pp. 48-52, 2009.

[20] J. Crain, L. Opyrchal, and A. Prakash, "Fighting phishing with trusted email", ARES '10, pp. 462-467, 2010.

[21] J. Zhang, W. Xu, Y. Peng, and J. Xu, "MailTrust: a mail reputation mechanism based on improved TrustGuard", CMC10, pp. 218-222, 2010.

[22] J. Bi, J. Wu, and W. Zhang, "A trust and reputation based anti-spim method", IEEE INFOCOM 2008, pp. 2485-2493, 2008.

[23] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web". Technical Report, Stanford University, 1998.

[24] Y. T. Liu, B. Gao, T. Y. Liu, Y. Zhang, Z. M. Ma, S. Y. He, and H. Li, "BrowseRank: letting web users vote for page importance", In *Proc. of SIGIR*. pp. 451-458, 2008

[25] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen, "Combating web spam with TrustRank", In *Proc. of VLDB*, pp. 576-587, 2004.

[26] A. Schwartz, SpamAssassin. O'Reilly Media, Inc., 2004.

[27] T. A. Meyer and B. Whateley, "Spambayes: effective opensource, bayesian based, email classification system," in CEAS, 2004.

[28] M. Wong and W. Schlitt, "Sender policy framework (SPF) for authorizing use of domains in e-mail, version 1", RFC 4408 (Experimental), April 2006. http://www.ietf.org/rfc/rfc4408.txt

[29] E. Allman, J. Callas, M. Delany, M. Libbey, J. Fenton, and M. Thomas, "DomainKeys identified mail (DKIM) signatures," RFC 4871 (Proposed Standard), May 2007. http://www.ietf.org/rfc/rfc4871.txt

[30] R. Haskins and D. Nielsen, "Slamming spam: a guide for system administrators. Addison-Wesley Professional, 2004.

[31] T. Grandison and M. Sloman, "A survey of trust in internet applications", IEEE Communications and Survey, vol. 3, no. 4, pp. 2-16, 2000.

[32] http://www.re2ee.org/