Unwanted Traffic Control via Hybrid Trust Management

Zheng Yan
Department of ComNet, Aalto University, Espoo, Finland
The State Key Lab of ISN, Xidian University, China
zheng.yan@aalto.fi

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

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

***Abstract*— At the same time as the Internet provides a lot of social value, it is bogged down by unwanted traffic, which is malicious, harmful or unexpected for its receiver. This paper proposes an unwanted traffic control solution through hybrid trust management. It can control unwanted traffic from its source to destinations according to trust evaluation at a Global Trust Operator and traffic and behavior analysis at hosts. Thus, it can support unwanted traffic control in both a distributed and centralized manner and in both a defensive and offensive way. Simulation based evaluation shows that the solution is effective with regard to botnet intrusion, malicious attack of ISP and DDoS intrusion via reflectors.**

Keywords- spam filtering; trust; trust managemen; trust model; reputation; malware detection.

I. INTRODUCTION

At the same time as the Internet provides a lot of social value, it is bogged down by unwanted traffic, which is malicious, harmful or unexpected for its receiver, e.g., spam, DDoS intrusion, malware, botnet intrusion, malicious attack and unexpected advertisement contents. Botnets are the major security threat in the Internet. They are used to spread malware, send spam, attack hosts and networks, collect sensitive information from users and earn money from fraud.

Fighting bots and botnets is difficult due to many technical and social reasons. On the technical plane, the person in command, i.e., the botmaster hides behind multiple layers of bots. On the social plane, security issues are difficult for ordinary users to comprehend leading to low security awareness. Thus, it is preferred to have an automatic and intelligent solution with minimum involvement of the users.

In our previous work, we propose a generic unwanted traffic control solution through global trust management [2]. It can control unwanted traffic from its source to destinations according to trust evaluation. We propose to build 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 the trust of each system entity in order to decide how to control the unwanted traffic from a specific source. The trust of an entity contains two parts: the global trust that indicates the probability and nature of unwanted traffic sourced from the entity and the detection trust that specifies the detection performance of each entity. The global trust management system adopts a centralized architecture with an assumption that the unwanted traffic can be detected at the host either manually or automatically with the support of installed toolkits. The toolkit is capable of detecting intrusions (e.g., Distributed Denial of Services-DDoS) targeting at a specific host. We also assume that each ISP timely and honestly forwards the reports from its hosts to the GTO in a secure way.

In this paper, we extend our previous work by introducing a hybrid trust management system to control unwanted traffic in both a distributed and centralized manner. Thus, the above two assumptions can be released towards practical system deployment. Concretely, except for evaluating each entity's global trust at GTO in order to figure out if the traffic from it should be controlled for a receiver, the host itself is capable of blocking traffic targeting on it based on local traffic and behavior analysis. We define that a counter approach to unwanted traffic is defensive if it is focused on protecting hosts and networks from the unwanted traffic using traffic and content analysis and blocking it based on local knowledge. The approach is offensive if it seeks to control unwanted traffic as well as punish malicious or indifferent behaviors and encourage good behaviors of hosts and ISPs. Therefore, the proposed system can filter unwanted traffic at each host in a defensive way and automatically control traffic from a distrusted source in an offensive manner. Although a number of trust and reputation mechanisms have been proposed for controlling spam [3-10], spim (i.e., Instant Messaging spam) [11], SPIT (Spam over Internet Telephony) [12] and web pages [13-16], to our knowledge, such a comprehensive solution as what we develop in this paper is still lacking in the previous work.

The rest of the paper is organized as follows. Section 2 gives a brief overview of related work. Section 3 introduces a hybrid trust management system structure followed by a procedure to comprehensively control unwanted traffic. The algorithms used in the system are described in Section 4. In Section 5, we evaluate the effectiveness of the proposed system through simulations by testing its performance under a number

of typical attacks. Finally, conclusions and future work are presented in the last section.

II. Related Work

A. Unwanted Traffic Control via Trust Management

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 [6]. 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. The design differs from our work in three folds: 1) a distributed trust management framework could cause extra traffic and processing loads with regard to trust information request and propagation. Our system adopts GTO to manage trust in order to reduce such a cost; 2) trust information is used to tune filter threshold, while we directly use the global trust to indicate if a traffic from a source should be controlled; 3) we apply the detection trust to tailor the considerations of evidence collected from different entities for global trust evaluation in order to overcome a number of potential attacks [2, 17].

Some existing spam control solutions cannot provide counter ways in both a defensive and offensive way. For example, a layered trust management framework was proposed in order to help email receivers eliminate their unwitting trust and provide them with accountability support [5]. IPGroupRep clusters the senders into different groups based on their IP addresses and computes the reputation value of each group according to the feedback of email receivers on the messages sent from the group [7]. The reputation value can be used to indicate whether an incoming message is spam or not. However, the above solutions cannot provide defensive protection at hosts and overcome such attacks as wrong/malicious feedbacks from ISPs.

Other spam control solutions adopt different system structure or mechanisms from our solution, although some features are similar to ours, e.g., MailTrust [10], spammer detection based on the behavior of email senders [8], and a multi-level reputation-based greylisting solution [9]. Comparing to the above work, the trust evaluation in our solution is not only based on the traffic and behavior analysis at hosts, but also the monitored behaviors of unwanted traffic sources at ISPs.

Highly related to our work, a reporter-based reputation system for spam filtering was proposed to filter spam [3]. The system includes a trust-maintenance component, in which users gain and lose reputation, depending on their spam-reporting patterns. The filtering component uses the reports of highly reputable reporters for spam removal, while in our solution the traffic control is based on all collected reports with the detection trust as a discount. This work didn't evaluate the system performance in various situations, such as DDoS intrusion via reflectors and attacks raised by malicious ISPs. The authors did not discuss its applicability on other kinds of unwanted traffic.

A number of solutions attempted to overcome web page spam [13-16], VoIP spam calls (SPIT) [12], spam of instant messaging (SPIM) [11]. These solutions are only applicable for a specific type of spam, not generic and suitable for controlling other types of unwanted traffic. Literature still lacks a comprehensive unwanted traffic control solution, which is efficient, accurate, and robust to control various types of unwanted traffic in both a distributed and centralized manner and in both a defensive and offensive way. Our solution proposed in this paper aims to solve this issue.

B. Global Trust Management vs. Hybrid Trust Management

The hybrid trust management solution provides a framework that has potential to control unwanted traffic in a comprehensive manner. It differs from our previous global trust management solution in the following aspects [2]:

1) Each host is capable of defending against unwanted traffic through analysis of inbound traffic and host behaviors;

2) Each host can request its local ISP or GTO to control unwanted traffic based on its personal analysis, thus the new solution supports personalized unwanted traffic control raised by hosts. GTO can also control unwanted traffic based on trust evaluation and past reporting behaviors of hosts;

3) Except for keeping the robustness against the attacks on the trust management system [2], the hybrid solution is capable of fighting against new traffic intrusion models and system attack models.

Our previous solution proposed in [2] can control various types of unwanted traffic in a centralized manner with an offensive way. We have evaluated its robustness over a number of malicious system attacks raised by hosts. In this paper, we extend our previous solution in order to provide a comprehensive unwanted traffic control solution, which is efficient, accurate, and robust to control various types of unwanted traffic in both a distributed and centralized manner and in both a defensive and offensive way. Meanwhile, we test its performance with regard to a couple of new unwanted traffic intrusion models, which are not defended by the previous one. Except for keeping the advance of the previous solution, the solution proposed in this paper performs well under malicious ISP attack, which is not supported by the previous one.

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 [18]:

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. Meanwhile, each content/traffic flow (i.e., a

sequence of packets from a source to a destination) can be identified based on its hash code.

2. GTO assumption: A Global Trust Operator behaves as an authorized trusted 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 some entities (e.g., other hosts) in the network.
4. Tracking assumption: We assume that the unwanted traffic source can be tracked based on analyzing traffic logs. For scalability we use trust management to control the logging.

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

(1) Timely/efficient and accurate defense against unwanted traffic intrusion at hosts;
(2) Efficient recognition of unwanted traffic sources under Botnet intrusion and DDoS intrusion;
(3) Automatic maintenance of trust for each system entity;
(4) Robustness against attacks raised by malicious ISPs.

B. Attack Model

In our previous work, we proved that the global trust management system is accurate to control unwanted traffic based on trust evaluation. It is efficient in controlling unwanted traffic caused by normal botnet infection. It can also overcome a number of system attacks raised by malicious hosts, such as hide evidence attack, on-off attack and bad-mouthing attack. In this paper, we focus on evaluating the accuracy and efficiency of the system under two kinds of unwanted traffic intrusion models:

- Extreme botnet infection: quite a number hosts are infected in the Internet, thus they attempt to send unwanted traffic to a limited number of hosts as their targets.
- DDoS via reflectors [19]: unwanted traffic could intrude one victim host from a number of attacked innocent hosts (reflectors). The unwanted traffic could be the same or different from different reflectors.

We further test the robustness of the system under malicious ISP attack as described below:

- Malicious attack of ISP: an ISP could maliciously perform an attack on the designed system. It behaves well to get a high trust value. It then turns its resources against the system. The malicious ISP could conduct a hide evidence attack by blocking all detection reports of its hosts or a bad mouthing attack by framing a good traffic source.

C. System Structure

Fig.1 shows the structure of the hybrid trust management system for unwanted traffic control. Differently from the global trust management, each host has a User Behavior Monitor (UBMo) to track the host behaviors with regard to unwanted traffic processing. A Local Traffic Monitor (LTMo) is applied to monitor inbound traffic to detect potential intrusions. The host also embeds an Unwanted Traffic Detector (UnTD), which can analyze the input data from UBMo and LTMo, as well as any unwanted traffic detection toolkits for detecting different kinds of spam or intrusions. An Evidence Reporter (EvR) at the host reports the unwanted traffic detection results to its local ISP.

At ISP, an Evidence Collector (EvC) collects the reports. A trust manager (TM) 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 analysis based on the evidence collected from local hosts and/or the input from LUMo. Both analysis results of the local ISP and GTO are used to trigger traffic monitoring at ISP (LUMo) and traffic similarity check. An ISP Trust Manager (IspTM) is responsible for transferring the results from LTM to GTO, requesting GTO for trust evaluation and unwanted traffic control, and receiving the trust value of the requested entity and a blacklist, as well as personalized traffic control decision from GTO.

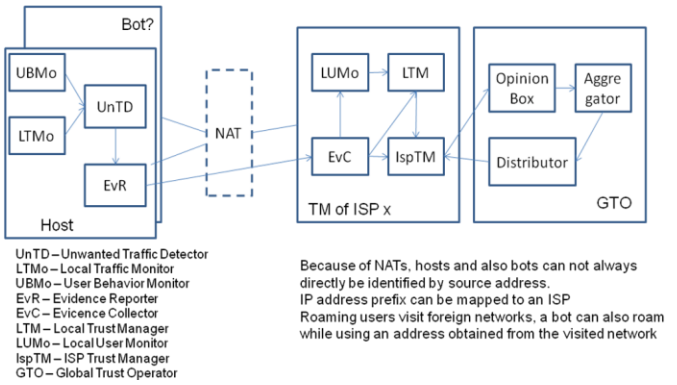

Figure 1. System structure of hybrid trust management

At the GTO side, an Opinion Box is used to securely store trust evidence and information that are used to evaluate the global trust and detection trust of each entity and make an unwanted traffic control decision at an Aggregator. The Aggregator can also map a traffic source to its ISP. At GTO, a Distributor is applied to collect trust evidence and information, receive requests from ISPs and distribute the commands and decisions of GTO to ISPs.

IV. UNWANTED TRAFFIC CONTROL PROCEDURE

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

Concretely, the host device monitors inbound traffic. If the monitored traffic flow is suspicious, the unwanted traffic process behavior of the correspondent host is further tracked and inbound traffic similarity is calculated in order to generate an unwanted traffic detection report. The host reports to its local ISP if the detection is positive. The ISP collects the complaint reports from hosts and forwards them to GTO. If the complaint on a local host is serious or GTO triggers, ISP does traffic monitoring on the suspicious entities and checks

content similarity. If the ISP checks are abnormal or suspicious, ISP sends its own check results to GTO. The GTO collects the reports from ISPs and hosts and then evaluates the global trust and detection trust of each system entity in order to detect the source of unwanted traffic and send a blacklist to ISPs. It requests to track the suspicious remote source of unwanted traffic by analyzing the reports from hosts and sends a command to the suspected attacker's local ISP to trigger traffic monitoring and similarity check. For personalized unwanted traffic control, the ISP sends a request to the GTO if it discovers some traffic sourced from a host in the blacklist. Based on the past detection reporting behaviors, the GTO will make a control decision for a particular destination. This mechanism is useful for filtering unwanted traffic that is not malicious but unexpected, such as advertisements. The system also supports controlling traffic for a specific host or ISP if it requests a personalized control. In addition, GTO can generate a personalized blacklist for each host and disseminate it to the local ISP of the host if personalized unwanted traffic control is needed.

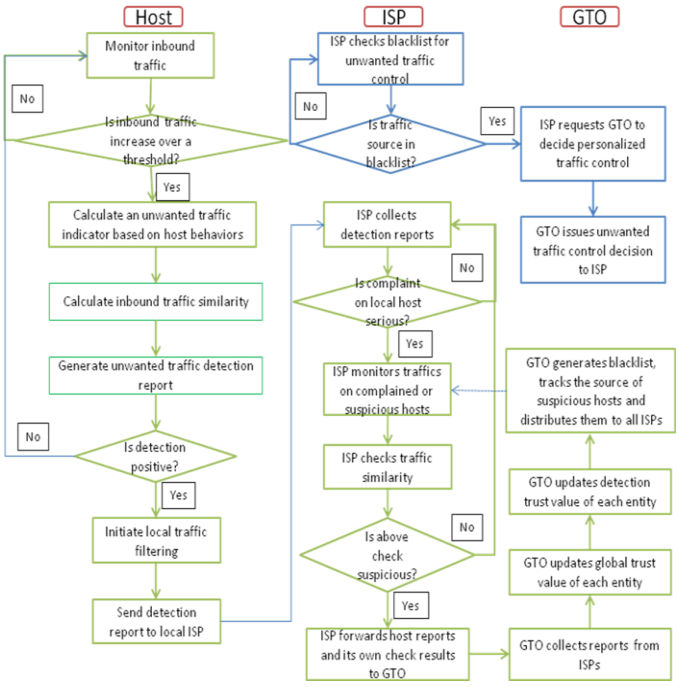

Figure 2. An unwanted traffic control procedure

V. ALGORITHMS

Based on the above system design and the unwanted traffic control procedure, we propose a number of algorithms to implement unwanted traffic control. For ease of reference, Table 1 summarizes the notations used in section V.

TABLE I. NOTATIONS OF ALGORITHM 1

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;
$tr_k^{in}(t)$	The inbound traffic flow 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*;
φ^k	The unwanted traffic indicator contributed by local traffic monitoring at host U_k;
e_i^k	The *i*th content received by host U_k;
r_t^i	The receiving time of e_i^k at U_k;
d_t^i	The deleting time of e_i^k at U_k;
τ_i	The unwanted traffic indicator contributed by the unwanted traffic process behaviors of a host regarding the *i*th content;
T	The time window used to normalize the unwanted traffic process time;
$v_k^i(t)$	The probability of e_i^k being an unwanted content indicated by U_k at time *t*, the unwanted traffic intrusion indicator;
$s_i^k(t)$	The unwanted traffic detection result at time *t* by U_k about e_i^k;
$s_i(t)$	The unwanted traffic detection result at time *t* about e_i^k;
$sim\_in_i^k$	The similarity of inbound traffic correlated to e_i^k;
sim_in^k	The similarity of U_k inbound traffic by considering all similar traffic received by U_k;
$\theta(I)$	The Rayleigh cumulative distribution function $\theta(I)=\left\{1-\exp(\frac{-I^2}{2\sigma^2})\right\}$ to model the impact of integer number I, $\sigma=100$ in our simulation;
$tr_k^o(t)$	The outbound traffic flow of U_k at time *t*;
ut_k^t	The global trust of U_k at time *t*;
thr	The threshold of the host to report to ISP;
*thr*0	The threshold to trigger traffic monitoring at local ISP;
*thr*1	The threshold of ISP to report to GTO;
φ_{sp}^k	The unwanted traffic indicator contributed by the ISP traffic monitoring on U_k;
$sim\_out_i^k$	The similarity of outbound traffic of U_k correlated to e_i^k;
sim_out^k	The traffic similarity factor of U_k by considering all similar traffic sent by U_k;
$sp_k^n(t)$	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 value of U_k at time *t*;
y	The detection performance indicator;
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;
*thr*2	The threshold to put an entity into the blacklist at GTO;
*thr*3	The threshold to determine on-off or conflict behavior attack;
*thr*4	The threshold to determine dishonest ISP.

A. Unwanted Traffic Detection at Host

1) Local Traffic Monitoring

The purpose to monitor the inbound traffic flow of a host U_k ($k = 1,.....,K$) at local device is to detect whether there is an attempt to intrude the host. For U_k, an unwanted traffic indicator contributed by the local traffic monitoring can be described as:

$$\varphi^k = \left|1 - 2f\left\{d_t\left\{tr_k^{in}(t)\right\}\right\}\right|. \quad (1)$$

2) Traffic Process

If the receiving time of a content e_i^k is r_t^i and its deleting time (or the time to move it to the spam folder) is d_t^i, an unwanted traffic indicator τ_i contributed by host behavior monitoring can be described as

$$\tau_i = 1 - \frac{d_t^i - r_t^i}{T}, \quad (2)$$

with an average value $\tau = \frac{1}{I}\sum_{i=1}^{I}\left(1 - \frac{d_t^i - r_t^i}{T}\right)$.

3) Similarity Check

If $\varphi^k \geq thr1$, we further check the similarity of contents received by U_k. For similar sized contents that have similar lengths or same hash codes) $E_k = \{e_i^k\}$ $i = \{1,.....I\}$ received by U_k within a time window ($w = \left[t - T/2, t + T/2\right]$), we calculate their similarity as:

$$sim\_in_i^k = \frac{\theta(I)}{I-1}\sum_{i' \neq i}^{I}\left(1 - \left|e_i^k - e_{i'}^k\right|\right), \quad (3)$$

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. Obviously, U_k could receive multiple sets of similar traffic intrusion. The similarity of U_k inbound traffic by considering all similar contents is

$$sim\_in^k = \frac{1}{M'}\sum_{M'}\left\{\frac{\theta(I)}{I-1}\sum_{i' \neq i}^{I}\left(1 - \left|e_i^k - e_{i'}^k\right|\right)\right\}. \quad (4)$$

where M' is the number of the sets of similar contents. In formula (3) and (4), we consider the influence of parameter *I* using Rayleigh cumulative distribution function $\theta(I)$.

4) Unwanted Traffic Reporting

A host could complain about unwanted traffic to its local ISP. The unwanted traffic detection value $v_k^i(t)$ at time *t* by U_k about traffic e_i^k is described as:

$$v_k^i(t) = sim\_in_i^k * \varphi^k * \tau_i. \quad (5)$$

The detection reports are aggregated at ISP in order to decide whether traffic monitoring and check at ISP is needed for a local traffic source. Aggregation is also conducted at GTO for a remote traffic source in order to decide whether traffic monitoring and check at its ISP is needed. The aggregation is based on Formula (6).

$$s_i(t) = \frac{\sum_k v_k^i(t) * ut_k^t * dt_k^t}{\sum_k ut_k^t * dt_k^t} \quad (6)$$

An unwanted traffic detection report containing $v_k^i(t)$ is automatically sent to the local ISP of the host if $v_k^i(t) \geq thr$.

Algorithm 1 is applied to detect and control unwanted traffic at a host.

Algorithm 1: Unwanted Traffic Detection and Control at a Host

1. Input:
2. - $tr_k^{in}(t)$, e_i^k, r_t^i, d_t^i ($i = 1,...,I$).
3. Monitor U_k's inbound traffic to get φ^k if the traffic is increasing;
4. **For** each suspicious content e_i^k, **do**
5. Calculate τ_i;
6. **If** $\varphi^k \geq thr1$, calculate $sim\_in^k(t)$ and $v_k^i(t)$;
7. **If** $v_k^i(t) \geq thr$,
8. Send $v_k^i(t)$ to local ISP, initiate local traffic filtering;
9. Filter the content from the same source or the similar contents from different sources.
10. Output: $v_k^i(t)$, ($i = 1,.....I_{U_k}$).

B. Traffic Monitoring at ISP

The purpose to monitor a host U_k's traffic at its local ISP is to find the source 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 source. This traffic monitoring is triggered by a condition $s_i(t) \geq thr0$ in order to save the running cost of ISP. Particularly, it can detect an infected host that has become 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. It is most efficient to monitor own subscribers because the ISP sees all traffic sourced at its own subscribers while other ISP's subscribers are numerous and the ISP can see only a fraction of their traffic. Therefore, for scalability, monitoring of other ISP's subscribers should be very selective. An unwanted traffic indicator contributed by the ISP traffic monitoring on the outbound traffic of U_k is

$$\varphi_{sp}^k(t) = \left|1 - 2f\left\{d_t\left\{tr_k^o(t)\right\}\right\}\right|. \quad (7)$$

Similarly, the similarity of multiple *M* different unwanted contents sent out from U_k can be described as:

$$sim\_out^k = \frac{1}{M}\sum_{M}\left\{\frac{\theta(I)}{I-1}\sum_{i'\neq i}^{I}\left(1-\left|e_i^k - e_{i'}^k\right|\right)\right\}. \quad (8)$$

The unwanted traffic detection value about U_k provided by the nth ISP at time *t* can be described as:

$$sp_k^n(t) = \varphi_{sp}^k(t) * sim\_out^k. \quad (9)$$

ISP reports its monitoring result $sp_k^n(t)$ to GTO if $\varphi_{sp}^k \geq thr1$ or $sp_k^n(t) \geq thr1$. Algorithm 2 is applied to monitor unwanted traffic at ISP.

Algorithm 2: Unwanted Traffic Monitor at ISP

1. Input:
2. - $tr_k^o(t)$, $E_k = \{e_i^k\}$ $i = \{1,.....,I\}$;
3. - U_k ($k = 1,...,K$).
4. **For** each complained U_k, **do**
5. Monitor U_k's traffic, calculate $\varphi_{sp}^k(t)$;
6. Calculate sim_out^k and $sp_k^n(t)$;
7. **if** $\varphi_{sp}^k \geq thr1$ or $sp_k^n(t) \geq thr1$
8. Report $sp_k^n(t)$ to GTO.
9. Output: $sp_k^n(t)$, $n = (1,...,N)$.

C. Unwanted Traffic Control at GTO

The GTO evaluates the trust of each entity based on collected reports from the hosts and ISPs in order to find the source of unwanted traffic. For each system entity $U_{k'}$, we aggregate the reports from *K1* hosts who blamed this source as below:

$$rt_{k'}^{t_p} = \frac{\sum_{k=1}^{K1} dt_k^{t_p} * ut_k^{t_p} * v_k^i(t) * e^{-\frac{|t-t_p|^2}{\tau}}}{\sum_{k=1}^{K1} dt_k^{t_p} * ut_k^{t_p} * e^{-\frac{|t-t_p|^2}{\tau}}}, \quad (10)$$

where t_p is the trust evaluation time, τ is a parameter to control the time decaying, ($\tau = 2$ in our simulations).

We further aggregate the reports from ISPs to calculate their contributions on $U_{k'}$'s global trust evaluation.

$$mt_{k'}^{t_p} = \frac{\sum_{n=1}^{N} dt_n^{t_p} * ut_n^{t_p} * sp_{k'}^n(t_p)}{\sum_{n=1}^{N} dt_n^{t_p} * ut_n^{t_p}}. \quad (11)$$

We evaluate the global trust value of the blamed entity *k'* by considering the number of blamers as:

$$ut_{k'}^{t_p} = ut_{k'}^{t_p} - \theta(K1) * rt_{k'}^{t_p} - \theta(N) * mt_{k'}^{t_p}. \quad (12)$$

Algorithm 3 is used to conduct global trust evaluation and unwanted traffic control.

Algorithm 3: 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 blamed system entity k', **do**
4. Calculate $rt_{k'}^{t_p}$, $mt_{k'}^{t_p}$, and $ut_{k'}^{t_p}$ based on (1)-(5), (7)-(12);
5. **If** $ut_{k'}^{t_p} \leq thr2$, put $U_{k'}$ into blacklist.
6. Output: blacklist $\{U_{k'}\}$.

D. Detection Trust: The Credibility of Detection

We introduce detection trust to indicate the credibility of unwanted traffic detection, which is another dimension of trust. The detection trust of U_k is generated at GTO. If the detection reported by U_k doesn't match the final evaluation result, $y = -1$, and $\gamma{+}{+}$; If the detection matches the fact, $y = 1$ and γ is not changed; If no detection report is provided, $y = 0$ and γ is not changed. The detection trust dt_k^t of U_k at time t is:

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

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.

VI. Evaluation and Analysis

A. Simulation Settings and Evaluation Measure

We design a number of simulations to evaluate the feasibility and effectiveness of the proposed system. This is because real data based evaluation is not suitable for testing the system performance under various intrusions and system attacks. 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. Each unwanted traffic source randomly selects a number of hosts to intrude. A good host which has not been infected reports unwanted traffic honestly and timely. A malicious or indifferent host reports the unwanted traffic with a malicious or indifferent pattern. In our previous work, we have evaluated the robustness of GTO to fight against the attacks raised by malicious or indifferent hosts [2]. In this paper, we examine the effectiveness of the system under DDoS intrusion via reflectors and botnet intrusion, as well as malicious attack of ISP. In our simulations, we assume that the unwanted traffic from the same source is 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 Parameters

Symbol	Settings	Symbol	Settings

thr	0.8	σ	100
$thr0$	0.7	τ	2
$thr1$	0.8	δ	0.05
$thr2$	0.1	μ	0.1
$thr3$	5	thr4	0.1

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 are sources of unwanted traffic (SUT) and are indeed detected as SUT as x; the number of entities that are not SUT but are added to SUT as y; the number of entities that are SUT but are not detected as SUT as z. With these values we do a precision-recall evaluation. We define:

$$R = \frac{x}{x+z}, \tag{15}$$

$$P = \frac{x}{x+y}, \tag{16}$$

$$F = \frac{2PR}{P+R}, \tag{17}$$

where $R, P, F \in [0,1]$. R indicates the performance of false negative detection (i.e., unwanted traffic goes unnoticed). P indicates the performance of false positive detection (i.e., the blame of innocent hosts). Good system performance requests both high recall R and high precision P. Thus, we make use of F measure to indicate the system performance. Obviously, High F measure is desirable for a good performance of the system.

B. *Experiment 1: Efficiency of Unwanted Traffic Source Detection against a Botnet Attack*

We design Experiment 2 to test the efficiency of unwanted traffic detection against an extreme botnet infection: 800 hosts (botnet) in 4 ISPs intrude 100 hosts in the 5th ISP, i.e., they send unwanted traffic to 100 hosts in the 5th ISP. At each time slot, each of these 800 botnet hosts randomly selects 100 hosts in the 5th ISP to intrude by sending the same content. We apply traffic function $tr_k^i(t) = \alpha t$, $\alpha =$ 10, 10^2, 10^3, 10^4, or 10^5 in this test.

Efficiency can be reflected by detection speed/performance, i.e., how fast the system can detect the sources of unwanted traffic. In this experiment, we test the F measure in the above simulation settings, and also show intrusion indication at the host device based on Algorithm 1.

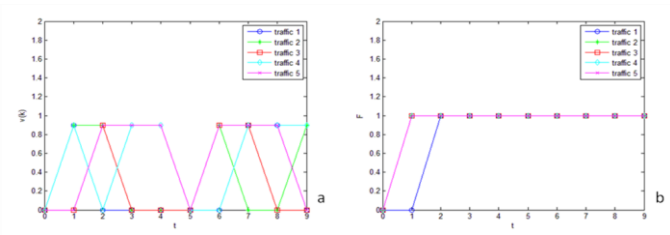

Figure 4. The efficiency of unwanted traffic detection in Botnet intrusion: a) unwanted traffic indication at a host; b) F-measure.

Fig.4 shows the result. We observe that a host can detect this kind of intrusion immediately. In some time slots, $v_k^i(t) = 0$, indicating that the host is not intruded by the unwanted traffic at those slots. This is because the target hosts are randomly selected. We also observe that the system can detect all unwanted traffic sources efficiently. The bigger volume the traffic is, the faster the detection. Note that in this experiment the hosts do not bad mouth evidence.

C. *Experiment 2: Efficiency of Unwanted Traffic Source Detection under a Malicious Attack of ISP*

In this experiment, we assume that the first ISP performs a malicious attack on the designed system. It behaves well and gets a high trust value 1, then conduct a hide evidence attack by not forwarding any detection reports from its hosts to GTO at the 10th time slot. Meanwhile, it performs a bad mouthing attack by framing a good traffic source as unwanted one. We apply traffic function $tr_k^i(t) = \alpha t$, $\alpha =$ 10.

Fig.5 shows the result. We observe from Fig.5.a that detection trust value of this malicious ISP is initiated at 0.5, then increased to 1 due to good behaviors, but dropped to 0 sharply at the 10th time slot since it is very easy for GTO to find this malicious ISP. In addition, the system can find all unwanted traffic sources quickly even though the malicious ISP hides evidence from its local hosts and conducts bad mouthing attack at the 10th time slot, refer to Fig.5.b. The F measure is 0 at the 10th time slot because we clear the blacklist at that moment. But we notice that the system can immediately find all unwanted traffic sources even though one ISP becomes malicious. This result implies that the system can efficiently detect the malicious ISP and thus ignore its influence on the unwanted traffic control. Another test shows that performing only hide evidence attack by one ISP won't influence much on the system efficiency. Thereby, we conclude that our system performs very well if some ISP suddenly turns into malicious.

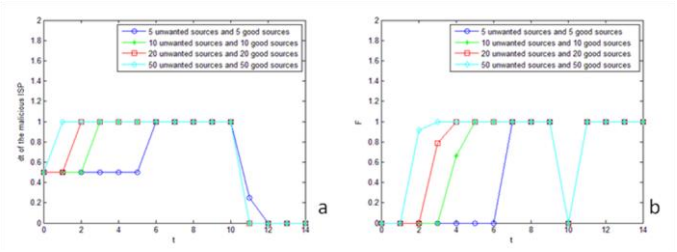

Figure 5. Unwanted traffic control performance under a malicious ISP attack: a) the detection trust of ISP; b) F measure.

D. *Experiment 3: Effectiveness of Unwanted Traffic Control for DDoS Instrusion via Reflectors*

We test two cases. In case 1, we randomly select N (N=100) hosts as reflectors that send the same contents to one target host. The simulation settings are: T = 10, $d_t^i - r_t^i = 1$, and M'=1. We test two traffic flows: (1) $tr_k^i(t) = \alpha t$, $\alpha =$ 10, 10^2, 10^3, 10^4, or 10^5; (2) $tr_k^i(t) = e^{\beta t}$, $\beta = 1$, 2, 4, 8, or 16. In case 2, the contents from N (N=100) reflectors are different,

but the contents from the same host are the same. That is M'=N. Other settings are the same as the case 1.

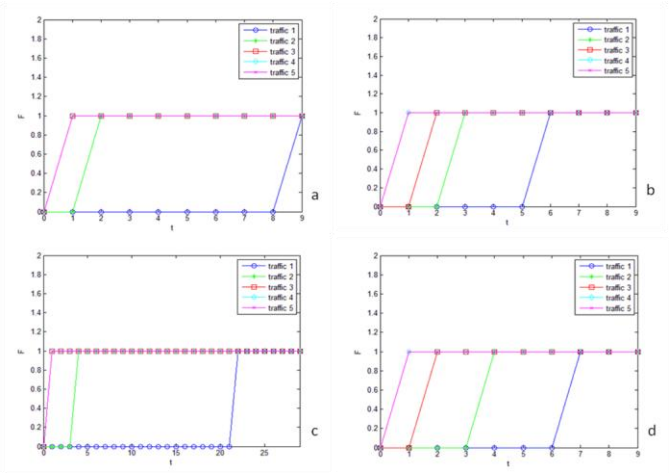

Figure 6. F measure of unwanted traffic detection in DDoS intrusion via reflectors: a) $tr_k^i(t) = \alpha t$, M'=1; b) $tr_k^i(t) = e^{\beta t}$, M'=1; c) $tr_k^i(t) = \alpha t$, M'=N; d) $tr_k^i(t) = e^{\beta t}$, M'=N. ($\alpha =$ 10, 10^2, 10^3, 10^4, or 10^5; $\beta = 1$, 2, 4, 8, or 16)

Fig.6 shows the result. We observe that the proposed system can detect all unwanted traffic sources within 9 time slots if all reflectors send the same contents to the target host. The system reacts slower in case 2 than case 1 when different reflectors send different contents to the target host, comparing Fig.6.a to Fig.6.c and Fig.6.b to Fig.6.d. This is reasonable since the volume of one content is smaller in the second case. The system is more sensitive when the volume of traffic is larger, refer to traffic 1-5 in each figure of Fig.6. This is because more detection reports are collected by GTO, thus it can find the unwanted traffic sources more efficiently by evaluating the global trust based on the detection reports.

In summary, the system performs accurately, efficiently and is robust under Botnet intrusion, DDoS intrusion via reflectors, and malicious ISP attack. The system performs more efficiently if the volume of traffic is larger It takes longer for the system to detect different unwanted contents than the same contents in the DDoS intrusion via reflectors.

VII. Conclusions and Future Work

This paper proposed an unwanted traffic control solution based on hybrid trust management by evaluating trust of each system entity at GTO and analyzing traffic and behaviors at hosts. We designed a number of algorithms that can be adopted by the system to control unwanted traffic in both a distributed and centralized manner and in both a defensive and offensive way. The simulation results show the effectiveness of our system with regard to accuracy and efficiency for unwanted traffic control, and robustness against system attacks raised by ISPs. Our paper contributes to the literature in two folds: (1) we proposed a hybrid unwanted traffic control framework over Internet, which is more comprehensive than previous work; (2) this system is effective to control unwanted traffic under different intrusion models, such as DDoS via reflectors and extreme botnet infection, and robust against attacks raised by a malicious ISP.

Regarding the future work, we will further improve the system by implementing it and investigating its performance in a real environment.

References

[1] OECD Broadband Portal, Bot infected computers, http://www.oecd.org/document/54/0,3746,en_2649_34225_38690102_1_1_1_1,00.html, referred Jan 25, 2012.

[2] Z. Yan, R. Kantola, Y. Shen, "Unwanted Traffic Control via Global Trust Management", IEEE TrustCom 2011, pp. 647 – 654, Changsha, China, Nov. 2011.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

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

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

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

[12] 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.

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

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

[15] 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.

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

[17] Y. Sun, Z. Han, and K. J. R. Liu, "Defense of trust management vulnerabilities in distributed networks," IEEE Communications Magazine, 46(2), pp.112-119, February 2008.

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

[19] M. Bossardt, T. Dubendorfer, and B. Plattner, "Enhanced Internet security by a distributed traffic control service based on traffic ownership", Journal of Network and Computer Applications, Vol. 30, No. 3., pp. 841-857, August 2007.