Monetizing Spambot Activity and Understanding its Relation with Spambot Traffic Features

Syed Fida Gillani
sgillan4@uncc.edu
University of North Carolina Charlotte

Ehab Al-Shaer
ealshaer@uncc.edu
University of North Carolina Charlotte

Sardar Ali
sardar.ali@nust.seecs.edu.pk
National University of Science and Technology

Syed Ali Khayam
ali.khayam@nust.seecs.edu.pk
National University of Science and Technology

Abstract

Spam botnets are no more driven by personal agenda, fun or proof of skills but by an underlying economic engine. Not until recently, intrusion detection techniques have approached spambot as a purely behavioral traffic detection problem using statistical features of mail traffic. Then, recently some efforts were made to comprehend the underly economic engine of spambot. These approaches either presents an abstract view of spambot economy or adapt purely measurement based approach to quantify spambot economy. No study so far has tried to bridge the gap between spambot detection and spambot economic modeling. We formalize the spambot economic system to monetize spammer efforts to spammer utility. We use standard consumer economic theory to translate spam activity to spammer utility. We also constrain this spammer utility through statistical features of mail traffic used by existing spambot detection techniques.

Keywords Botnet detection, spam economics, information theory, IDS tuning.

1. Introduction

A myriad of studies are reporting an exponential increase in the number and size of worldwide botnets [1, 2, 15, 17, 20, 20, 21]. For instance, it has been reported that the Storm botnet increased by a factor of three during the second quarter of 2008. The reason of such exponential growth is the financial gain that these spam botnets can generate [1–3]. Absent grounded empirical data, it is challenging to reconcile "revenue estimates" that can range from $2M/day for one spam botnet [4]. Paxson et. al [1] have documented 82,000 and 37,00 monthly orders for seven counterfeit pharmacies and counterfeit software stores, respectively. The spammers running all these spams generally purchase time from a bot master to launch a spam campaign with a single objective to increase their respective profit margins from such spam campaigns.

Until recently, most existing techniques, meant to block or filter spam activity, relied on statistical features of mail traffic [6, 11, 16, 18, 22, 23]. The effectiveness of all these techniques is measured in terms of high spam activity detection with low false positives, but is limited due to innate operational complexities and inherent uncertainties [7]. However, absent a rigorous treatment, the resulting information vacuum is all too easily filled with opinions, which in turn can morph into fact over time. However, while these same technical aspects were emphasized a lot by security community, recently, researchers [1, 2, 5, 10, 12] have explored the underlying economic engine that drives this ecosystem. Some of them [1, 2] have performed a rigorous measurement based studies of the financial aspect of spam activity and others [10, 12] abstracts the revenue model of spammers and bot masters. Thus far, however, no study has considered both economics and statistical features of mail traffic in relation to each other.

In this paper we make two contributions:

- We formalize the spam economic system to materialize spammer efforts into revenue.
- Constrained this financial spam ecosystem through statistical features of mail traffic.

In first case, we build a revenue/economic model to monetize the modern spam activity by adapting the general consumer theory of economics. In spam economy, spammer is a consumer looking for commodities and every commodity has an associated utility and like any other consumer spammer want to maximize his utility. We start by defining a commodity for spammer (consumer). To our best knowledge, we consider all relevant factors of interest, that a spammer may consider to rent a botnet, and formalize a commodity for spam economy. Then, come the next step that how a consumer will choose different commodities, which basically depicts the choice behavior of consumer. We have assumed a rational behavior of spammer, which simplifies the choice structure but for a rational behavior of consumer, commodity model must exhibit some certain properties. We show that our spam commodity model exhibit all those properties and it is safe to assume rational behavior. As a last step in economic modeling, we establish the utility function of a commodity to calculate the utility associated with each commodity. The objective function of spammer in our modeling is to maximize this utility.

In second case, we introduce statistical features of mail traffic as limiting forces in spam economic model to restrict the total utility. We consider following statistical features, as discussed by existing literature [6, 11, 16, 18, 22, 23]: inter-departure time, number of emails per recipients, distribution of new recipients per sampling window and size of spam email. It is claimed in the literature that these features can discern spam behavior from normal behavior. As a proof of concept we only constrain using inter-departure time and number of emails per recipients.

2. Related Work

Ramachandran and Feamster [16] focused on the network properties of spam and showed that network-level characteristics of spam are sufficiently different than those of legitimate emails. The work in [6] detects spam from email server logs by measuring changes over time from a particular source. In [22], a framework, called AutoRE, was presented to filter out any legitimate URLs and focus on the URL that the spammer wants his victims to click on to buy his merchandise or download his malware. Using their signature method, Xie et al. were able to identify botnet membership and determine which bots were used in the various spam campaigns.

Li et. al [12] proposed an abstract economic model for both both botmaster and attacker. They focused on the use of botnets for DDoS attack. After formalizing the model they introduce the concept of honeypots: fake bots which increase the probability of failure for attacker as he does not know how many bots he needs to launch a successful attack. However, the authors do not associate their model to any parameters used by their filters.

Table 1. Variables used in SPAM Economic Model

Notation	Definition
B_i	Average bandwidth of bots (bits/sec)
S	Minimum size of spam email (bits)
S_q	Average spam email size ($S \geq S$) (bits)
ν	Spam output rate of botnet
x	Quantity rented of output rate of botnet
c	Spam commodity (total botnet outcome $\nu * x$)
K	Number of commodities available
$\Re^K$	All possible consumption sets
β_l	Consumption bundle
C	Aggregated outcome of consumption bundle
W	Wealth of Spammer
P_l	Price associated with consumption bundle
$L_{(P,W)}$	Competitive budget set
$Pr\{c\}$	Probability of failure of spam commodity
$u(.)$	Spammer Utility Function

Steven et al. [11] proposed entropy and machine-learning detectors to differentiate between human and chat bots. A recent study on spamming botnet detection by Yao [23] proposed to detect bots (BotGraph) by constructing a large user-user graph. Kyle et al [18], used entropy to measure the effectiveness of different traffic features in identifying the spam behavior. Chris and Paxson et al [1, 2] are very prominent recent studies that have rigorously analyzed the spam economics using measurement based approach. They have established some very important statistics about spam financial ecosystem.

As mentioned earlier we are different from all these approaches because we are not just formalizing the spam economic system but to establishing a relationship between spam financial system and statistical features of mail traffic, and we also materialize this relationship.

3. Spam Botnet Economics

Spam ecosystem is being driven by underlying financial engine. It has become an economic system with all driving forces that are a part of any other economic system. From an economic perspective, every economy has two vital entities: *producers* (botmasters) and *consumers* (spammers). Botmasters want to reduce the production cost of the product (bots) and spammers want to reduce the buying cost of the product (renting botnets). We base our economic model upon the basics of consumer theory of economics and we only concern ourselves with the spammer not the botmaster. Similar to any other business, cost is the pivotal point of the spamming industry and therefore it is important to study economic issues when discussing shutting down spammers [1, 2, 5, 9, 19]. So, as first contribution we build an economic model for spammer that calculates the revenue associated with a spammer activity in terms of utility.

3.1 Economic Model Preliminaries

Economics is the study of the choices people make about commodities (products or services) as the result of scarcity. In spam economy, spammer is a consumer looking to choose

any commodity with maximum utility. We start by defining a commodity for spammer/consumer. To our best knowledge, we consider all relevant factors of interest, that a spammer may consider to rent a botnet, and formalize a commodity for spam economy. An exhaustive list of such factors in the present problem is: 1) Botnet size; 2) Bot bandwidth; 3) Spam mail template size; 4) Cost of victim email addresses; 5) Unit time cost of the rented botnet; 6) Maintenance botnet cost; 7) Mail content generation; 8) Active duration of botnet; 9) Response rate (number of mails get positive response from the end user.); and 10) Profit per response.

From a spammer's perspective, maintenance cost is already a part of the botnet rent cost, so it is no direct concern to the spammer. Cost of victim email addresses [9] and mail content generation cost are not recurring costs, because a spammer can always reuse the same victim email addresses or mail contents, and in economic theory a commodity cost has to be recurring in nature. Similarly, response rate and profit per response are not in control of the spammer or spamming industry and cannot impose any constraint on renting botnet decision. This phases out these factors from the final deciding factors that affect botnet selection decision. If we merge the remaining factors, botnet size, bot bandwidth and spam mail template size, we can derive another factor, referred to as *spam mail output rate*, as in Equation. 1:

$$\nu = \sum_{i=1}^{N} \frac{B_i}{S} \tag{1}$$

where i represent a bot, B_i represents the bandwidth of the i^{th} bot, S is the spam mail template size and botnet has N bots. We call this *spam mail output rate* (ν). When a spammer choose a commodity it chooses some desired quantity, let x be the quantity of a commodity ν. Then $c = x * \nu$ represents the total outcome (work) of the commodity[1]. This covers the active duration factor as mentioned above. We call c as basic commodity of our model. A spammer's choice of botnets is always restricted either because of botnet cost, botnet reliability, geographical constraints etc. So, in spam economy spammer has a finite number, say K number of available commodities. Normally in real, spammers rent multiple botnets to achieve some aggregated outcome. Let, β is a set comprises of any arbitrary combination of commodities that a spammer can choose. We call this a consumption bundle. As there are total K commodities, there are total $\Re^K$ possible consumption bundles. Every consumption bundle β_l (where $1 \leq l \leq \Re^K$) has an associated aggregated outcome C_l given by Equation. 2.

$$C_l = \sum_{\forall c_k \in \beta_l} c_k \tag{2}$$

[1] We use work and outcome inter-changeably in the paper.

Every commodity c_k has a price p_k and every consumption bundle β_l has as an associated price set, $P_l = \{\forall p_k | c_k \in \beta_l\}$. There exists some consumption bundles that a spammer cannot afford to choose due to his limited wealth (W). So, we call the set of all possible consumption bundles that a spammer can afford as the competitive budget set. This is formally defined in Equation. 3 as:

$$L_{(P,W)} = \{\forall\, \beta \in \Re^K | \beta_l^T \times P_l \leq W\}. \tag{3}$$

where β_l^T is simply the transpose of β_l.

This finalizes the definition of spam commodity. Then, come the next step that how a consumer will choose different commodities, which basically depicts the choice behavior of consumer. We have assumed a rational behavior of spammer, which simplifies the choice structure but for a rational behavior of a consumer, commodity model must exhibit some basic properties. We show that our spam commodity model exhibit all those properties and it is safe to assume rational behavior of spammer. This leads to the last step of economic modeling, establishing the spammer utility function. Because that will decided how a spammer compares different consumption bundles.

3.2 Formulation of a Spammer's Objective Function

The choice that a consumer makes is called a *Preference Relation* $\succeq$. It is a binary relation on the set of alternatives of consumption bundles, allowing the comparison of pairs of consumption bundles. If, β_1 and β_2 are two consumption bundles then $\beta_1 \succeq \beta_2$ means β_1 is *at least as good as* β_2 and $\beta_1 \succ \beta_2$ means β_1 is *preferred* to β_2. The role of the preference relation between consumption bundles is very critical as its absence makes an economic model unsolvable.

We assume that the choice behavior of spammer is rational, which demands certain properties to be true for the model. We discuss these as follows:

1. Preference relation $\succeq$ should be *rational* i.e. it should be complete and transitive. Completeness implies that $\forall \beta_l, \beta_j \in L_{(P,W)}$ either $\beta_l \succeq \beta_j$ or $\beta_j \succeq \beta_l$. Transitivity says that $\forall \beta_l, \beta_j, \beta_m \in L_{(P,W)}$, if $\beta_l \succeq \beta_j$ and $\beta_j \succeq \beta_m$ then $\beta_l \succeq \beta_m$. *Explanation:* Every commodity represents an outcome and every consumption bundle has an aggregated outcome, so all consumption bundles in competitive budget set hold this property.
2. Preference relation $\succeq$ should be *monotone*: if a consumption bundle β_l has more number of commodities than another consumption bundle β_j, then $\beta_l \succeq \beta_j$. *Explanation:* There may exist two consumption bundles with same aggregate rate, how to choose the preferred one? An earlier study [12] introduced the concept of virtual bots, thus to create an uncertainty about the success of a botnet outcome. They call it the probability of failure for each botnet and it is independent from each other. Let's assume that the probability of failure due to un-

certainty for each commodity is same, say $Pr\{c_k\}$. Suppose there exist two consumption bundles, $\beta_1 = \{c_1, c_2\}$ and $\beta_2 = \{c_3\}$ and $C_1 = C_2$. Then, the associated probability of failure of both consumption bundles are $Pr\{\beta_1\} = Pr\{c_1\} * Pr\{c_2\}$ and $Pr\{\beta_2\} = Pr\{c_3\}$. We use the same concept in our model and in this example β_1 is preferred over β_2. This satisfies the monotone property.

3. The preference relation, $\succeq$, should be *convex* such that for every $\beta_l \in L_{(P,W)}$ the upper contour set is convex: if $\beta_j \succeq \beta_l$, $\beta_m \succeq \beta_l$, and β_m not equal to β_j, then $\alpha\beta_j + (1-\alpha)\beta_m \succeq \beta_l$ for any $\alpha \in [0,1]$. *Explanation:* In standard economic theory, there are two reasons to impose this assumption: a) consumers typically like to consume mixed consumption bundles, i.e., it is better to use another consumption bundle with more number of commodities than one with single commodity; and b) It diminishes the marginal rate of substitution. Let us map both of these reasons to our spamming commodity structure. Reason (a) is catered for in second property above. For reason (b), suppose we have a function $F()$ that calculates the utility of a consumption bundle. If there is bundle $\beta_1 = \{c_1, c_2\}$ then by diminishing the marginal rate of substitution, given by $\frac{\partial F(\beta_1)}{\partial c1} / \frac{\partial F(\beta_1)}{\partial c2}$, the consumer requires more units of c_1 to remove one unit of c_2 to get the same utility. This gives stability to a consumption bundle by creating an area of indifference around it and indirectly gives a confidence to the spammer in his selection. Same holds in our commodity structure as the confidence value associated with each commodity and evasion due to diversity (second property) make the associated cost of each substitution non-linear.

The eventual goal of a spammer is to maximize the total utility, so we need to formalize an utility function for spam commodity. According to the definition of utility function:

DEFINITION 1. A function $u : \beta_l \rightarrow L_{(P,W)}$ is a utility function representing preference relation $\succeq$ if $\forall \beta_l, \beta_j \in L_{(P,W)}, \beta_l \succeq \beta_j \Leftrightarrow u(\beta_l) \geq u(\beta_j)$.

We formulate our utility function for each commodity considering three factors: 1) Commodity outcome; 2) Diversity of consumption bundle; 3) Failure probability of consumption bundle. Equation. 4 defines the utility function for a consumption bundle as:

$$u(\beta_l) = \sum_{\forall c_k \in \beta_l} c_k + (|\beta_l| * (1 - Pr\{c_k\})) \qquad (4)$$

where $|\beta_l|$ gives the count of commodities in a consumption bundle β_l. As mentioned earlier the goal of a spammer is to maximize his utility, than concludes the objective function in Equation. 5.

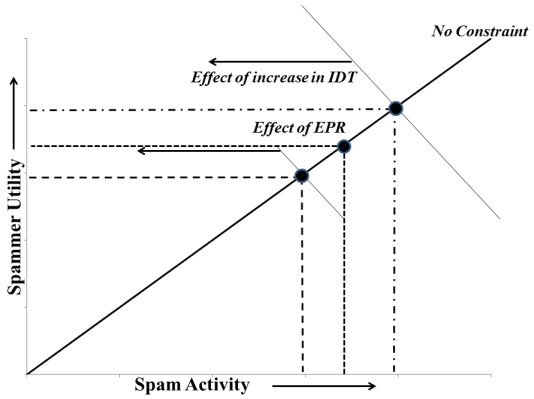

Figure 1. Abstract view of Spammer utility with and without IDT and EPR constraints.

$$\begin{aligned} &\max u(\beta_l) \\ &\textbf{\textit{s.t.}} \\ &\beta_l \in L_{(P,W)} \\ &\beta_l \neq \emptyset \\ &C_l > 0 \end{aligned} \qquad (5)$$

4. Constraining The Spam Economic Model

As mentioned before, the second objective of this paper is to constrain the spammer utility function using statistical features of mail traffic and materialize the impact of each feature. From existing literature [6, 11, 16, 18, 22, 23] we select critical traffic features and we discuss them as under.

4.1 Inter-Departure Time (IDT)

The foremost distinctive feature is the inter departure time between mails. It is the time between two consecutive emails. Spammers want to send as many emails as possible in a small time period to maximize the outcome. Let, Δt be the time interval between two consecutive spam mails. Even though it changes during the course of spam campaign but in end mean IDT ($\mu_{(idt)}$) is sufficient to calculate the volume of total outcome. A commodity c_k represents the total outcome of a botnet with $\mu_{(idt)} = 0$ and increasing the IDT really reduces the effective outcome of a commodity. So, the Equation. 1 is changed to Equation. 6.

$$\widetilde{\nu} = \sum_{i=1}^{N} \frac{B_i}{S + (\mu_{(idt)} * B_i)} \qquad (6)$$

The intuition of Equation. 6 is simple. As B_i represents the bits per second (unit time) and $\mu_{(idt)}$ represents a pause, which can be translated into the loss of bits that could have been sent otherwise. Figure. 1 shows the possible impact of this constraint on spammer utility..

4.2 Emails Per Recipient (EPR)

A normal user tend to send most of his mails to only a group of recipients, to whom he is connected socially or through business. On the other hand, a spammer avoids sending too many mails to a subset of recipients to avoid detection. Let, ω be the total number of recipients email addresses owned by spammer. If a spammer can safely send θ number of emails to each recipient without alerting any spam filter then the aggregated outcome is constrained by $\theta * \omega$, so ideally $\theta * \omega > \max(C_l \in L_{(P,W)})$. Even though spammer may have more wealth (W), but EPR can potentially further constrain the competitive budget set, see Figure. 1. So, the new competitive budget set constraint is given by Equation. 7.

$$\widetilde{L_{(P,W)}} = \{\forall\, \beta \in \Re^K | \beta_l^T \times P_l \leq W, C_l \leq \theta * \omega\}. \quad (7)$$

5. Conclusion

Spam botnets are no more driven by personal agenda but by the underlying economic engine. Most intrusion detection techniques had approached spam botnets as a purely behavioral traffic detection problem using statistical features of mail traffic. Recently some efforts were made to comprehend the underly economic engine of spam. These studies either took the road to provide abstract economic model or took a measurement based approach to quantify spam economy. In this paper we have formalized the spam economic system to monetize spammer efforts into utility. We have used standard consumer economic theory to calculate the spammer utility. We have also constrained our economic model using traffic features, inter-departure time and emails per recipients, discussed by existing literature as key features to discern spam traffic.

References

[1] Chris Kanich, Nicholas Weaver, Damon McCoy, Tristan Halvorson, Christian Kreibitch, Kirill Levchenko, Vern Paxson, Geoffrey M. Voelker, and Stefan Savage. *Show Me the Money: Characterizing Spam-advertised Revenue*. In *USENIX SECURITY SYMBOSIUM'11*, August 2011.

[2] Chris Kanich, Christian Kreibich, Kirill Levchenko, Brandon Enright, Geoffrey M. Voelker, Vern Paxson, and Stefan Savage. *Spamalytics: An empirical analysis of spam marketing conversion*. In *CCS'08*, October 2008.

[3] Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. *spam: The Underground on 140 Characters or Less*. In *CCS'10*, October 2010.

[4] C. Akass. *Storm Worm "Making Millions a Day"*. http://www.computeractive.co.uk/pcw/news/1923144/storm-worm-millions-day, 2008.

[5] Ania Monaco. *Cutting Down on Spam*. http://theinstitute.ieee.org/technology-focus/technology-topic/cutting-down-on-spam, October 7, 2011.

[6] Richard Clayton. *Stopping spam by extrusion detection*. *University of Cambridge*, 2004.

[7] Ben Laurie, and Richard Clayton. *"Proof-of-Work" Proves Not to Work*. In *WEIS04*, May 2004.

[8] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & sons, 1991.

[9] Organisation for Economic Co-Operation and Development. Malicious software (malware): A security threat to the internet economy. Ministerial Background Report DSTI/ICCP/REG(2007)5/FINAL, OECD, 2007.

[10] Richard Ford and Sarah Gordon. Cent, five cent, ten cent, dollar: Hitting botnets where it really hurts. In *ACM 2006, Session: Malware*, pages 3–10, August 2006.

[11] S. Gianvecchio, M. Xie, Z. Wu, and H. Wang. Measurement and classification of humans and bots in internet chat. In *Proceedings of the 17th conference on Security symposium*, pages 155–169. USENIX Association, 2008.

[12] Zhen Li, Qi Liao, and Aaron Striegel. Botnet eonomics: Uncertainty matters. In *Managing Information Risk and The Economics of Security*, pages 245–267, 2009.

[13] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[14] Ezekial Moving Ministry. What are the email sending limits of isps and other providers? http://support.e-zekiel.com/templates/Manual/details.asp?id=31606, Nov. 2007.

[15] Abhinav Pathak, Feng Qian, Y. Charlie Hu, Z. Morley Mao, and Supranamaya Ranjan. Botnet spam campaigns can be long lasting: evidence, implications, and analysis. In *SIGMETRICS '09: Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems*, pages 13–24, New York, NY, USA, 2009. ACM.

[16] Anirudh Ramachandran and Nick Feamster. Understanding the network-level behavior of spammers. In *SIGCOMM'06*, September 2006.

[17] Threat Research and Content Engineering. Srizbi now leads the spam pack. http://www.marshal.com/trace/traceitem.asp?article=567, Feb. 2008.

[18] Kyle Smith, Ehab Al-Shaer, and Khalid Elbadawi. Information theoretic approach for characterizing spam botnets based on traffic properties. In *IEEE ICC*, pages 1–5, 2009.

[19] Tim Wilson. Competition may be driving surge in botnets, spam. *DarkReading*, 2008.

[20] Staff Writers. Messagelabs: Storm botnet was 20% of spam. *Dark Reading*, 2008.

[21] Staff Writers. Spam up by 50% in first quarter of 2008. *SC Magazine*, 2008.

[22] Yinglian Xie, Fang Yu, Kannan Achan, Rina Panigrahy, Geoff Hulten, and Ivan Osipkov. Spamming botnets: Signatures and characteristics. In *SIGCOMM08*, August 2008.

[23] Y. Zhao, Y. Xie, F. Yu, Q. Ke, Y. Yu, Y. Chen, and E. Gillum. Botgraph: Large scale spamming botnet detection. In *Proceedings of the 6th USENIX symposium on Networked systems design and implementation*, pages 321–334. USENIX Association, 2009.