Identifying Spammers by Their Resource Usage Patterns

Kevin S. Xu
EECS Department
University of Michigan
1301 Beal Avenue
Ann Arbor, MI 48109-2122
xukevin@umich.edu

Mark Kliger
Medasense Biometrics Ltd.
PO Box 633
Ofakim, 87516 Israel
mark@medasense.com

Alfred O. Hero III
EECS Department
University of Michigan
1301 Beal Avenue
Ann Arbor, MI 48109-2122
hero@umich.edu

ABSTRACT

Most studies on spam thus far have focused on its content or source. These types of studies, however, reveal little about the behavioral characteristics of spammers. In addition, privacy issues may prevent wide access to email content. In this paper, we try to identify spammers by investigating their resource usage patterns. Specifically, we look at usage patterns of harvesters, the bots that are used to acquire email addresses, and spam servers, the email servers being used to send the spam emails. We perform spectral biclustering on both harvesters and servers to reveal groups of resources that are used together, which we believe correspond to individual spammers or groups of spammers. We make several interesting discoveries including a division into phishing and non-phishing spammers and a group of harvesters with highly correlated behavior that have IP addresses belonging to a known rogue Internet service provider.

1. INTRODUCTION

Previous studies on spam have mostly been focused on filtering methods based on the content of spam emails or blacklisting methods based on IP address reputation. In this paper, we take a different approach to studying spammers' behavior. Namely, we investigate spammers' resource usage patterns and try to identify spammers by finding groups of resources that are used together. The resources we investigate are *harvesters*, the bots that are used to acquire email addresses, and spam *servers*, the actual email servers that are used to send the spam emails. The harvester and server are the two intermediate steps in the path of spam, illustrated in Figure 1. Note that the spammer is unknown; we only observe his behaviors through his resource usage.

The source of the data analyzed in this paper is Project Honey Pot [2], a web-based network for monitoring harvesting and spamming activity by using trap email addresses. By distributing each trap address only once, we can uniquely identify the harvester that acquired each address. In a previous study [10], we assumed that harvesters are closely related to spammers and attempted to reveal groups of spammers by clustering harvesters. In this study, we cluster servers as well and identify spammers with biclusters of harvesters and servers.

We associate each email with the harvester that acquired the email address and the server that was used to send the email. In this manner, we construct a bipartite graph of harvesters and servers, where an edge is present between a harvester and server if at least one email is associated with both of them. We present a biclustering analysis that allows us to discover groups (biclusters) of harvesters and servers that are often used together. We believe that such biclusters correspond to individual spammers or groups of spammers who are working together by sharing resources. Since spammers are unobservable, we cannot distinguish between a single spammer and a group of spammers.

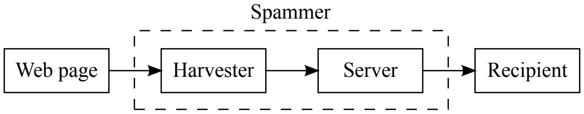

Figure 1: The path of spam from an email address on a web page to a recipient's inbox.

Project Honey Pot is an ideal data source for studying phishing emails, which attempt to fraudulently acquire sensitive information by appearing to represent a trustworthy entity. It is impossible for a trap email address to, for example, sign up for a PayPal account, so all emails supposedly from financial institutions can immediately be classified as phishing emails. We find that both harvesters and servers divide into dedicated phishing and non-phishing harvesters and servers, respectively.

Our main findings from the biclustering analysis are as follows:

1. Biclusters appear to split into phishing and non-phishing biclusters, indicating that most spammers are either phishing or non-phishing spammers. In addition, we find most phishing spammers do not share resources with non-phishing spammers and vice-versa, further suggesting that these two classes of spammers have little to no interaction.

2. Non-phishing biclusters form a core-periphery structure, suggesting that many non-phishing spammers send some spam emails using heavily shared or public resources. Conversely, phishing biclusters are more scattered, indicating that most phishing resources are private.

3. Several biclusters consist of harvesters or servers that have similar IP addresses, indicating physical proximity. In particular, we find a group of fifteen harvesters with the same /24 IP address prefix, which happens to be owned by a known rogue Internet service provider (ISP).

The first and third findings were also observed in [10] but for harvesters only; hence, this study both validates those findings and extends them to servers.

2. PROJECT HONEY POT

Project Honey Pot is a distributed system for monitoring harvesting and spamming activity via a network of decoy web pages

CEAS 2010 - Seventh annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference July 13-14, 2010, Redmond, Washington, US

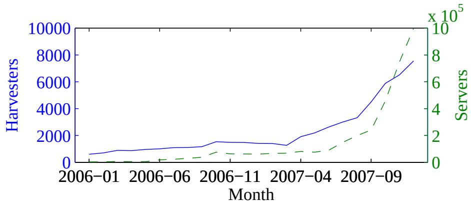

Figure 2: Number of active harvesters and servers by month.

with trap email addresses, known as *honey pots*. These trap addresses are embedded within the HTML source of a web page and are invisible to human visitors. As a result, harvesters tracked by Project Honey Pot consist solely of automated harvesting bots. While spammers have other means of acquiring email addresses, we consider only automated harvesters in this paper.

Each time a harvester visits a honey pot, the centralized Project Honey Pot server generates a unique trap email address. The harvester's IP address is recorded and sent to the Project Honey Pot server. The email address embedded into each honey pot is unique, so a particular address could only have been collected by the visitor to that particular honey pot. Thus, when an email is received at one of the trap addresses, we can associate it with a particular harvester. These addresses are not published anywhere besides the honey pot, so we can assume that all emails received at these addresses are spam. We refer readers to [2, 7, 10] for additional information on Project Honey Pot.

2.1 Harvester and Server Statistics

Project Honey Pot was founded in October 2004 and has grown exponentially with time. As of June 2010, over $78,000$ harvesters and 67 million servers are being monitored at over 45 million trap addresses [2]. The number of harvesters and servers monitored has grown exponentially with time. The extremely large number of servers tracked in 2008 and beyond currently prevents us from performing biclustering on more recent data, so we limit our analysis to the two-year period beginning in January 2006 and ending in December 2007. The number of active harvesters and servers in each month in this two-year period are shown in Figure 2.

As mentioned earlier, Project Honey Pot is an ideal data set for studying phishing emails. No legitimate emails from financial institutions would ever be received at the trap addresses, so emails containing phishing words can instantly be classified as phishing emails. The list of such words was built using common phishing words such as "password" and "account" and names of financial institutions that do business on-line such as PayPal.

For each harvester and server, we define a *phishing level* as the ratio of the number of phishing emails to the total number of emails associated with it. An interesting finding is that *most harvesters and servers are either associated only with phishing emails or only with non-phishing emails*. In particular, 62% of harvesters have a phishing level lower than 0.01, and 17% have a phishing level higher than 0.99. Also, 93% of servers have a phishing level lower than 0.01, and 3% have a phishing level higher than 0.99. We label a harvester or server as a phisher if its phishing level exceeds 0.5 and as a non-phisher otherwise. The labeling of harvesters as phishers or non-phishers will be used later when interpreting the biclustering results.

2.2 Bipartite Graph of Harvesters and Servers

Each email is associated with two entities, namely the harvester and server. We construct a bipartite graph of harvesters and servers, where an edge between harvester i and spam server j is present if at least one email is associated with both i and j. We choose the weight of the edge between i and j to be $h_{ij} = p_{ij}/e_i$, where p_{ij} denotes the number of emails sent by server j to addresses acquired by harvester i, and e_i denotes the total number of addresses acquired by harvester i. e_i is a normalization term to remove observation bias as a result of harvesters acquiring different numbers of addresses. This bias is partially due to the non-uniform distribution of honey pots on the Internet. Our study focuses on the structure of the bipartite graph, which we explore using biclustering.

3. BICLUSTERING METHODOLOGY

In traditional (one-way) clustering, data samples are partitioned into groups such that samples within a group are highly similar and samples between groups are highly dissimilar with respect to a set of features. Biclustering, also known as co-clustering or two-way clustering, differs from one-way clustering because it simultaneously partitions the samples and features. Biclustering is useful in applications where we are interested in the cluster structure of two classes of objects or attributes. Common applications include simultaneous clustering of words and documents [3] and simultaneous clustering of genes and conditions [4].

One-way clustering was performed on harvesters in [10]. In this paper we perform biclustering on both harvesters and servers. As mentioned in Section 2.2, we have a weighted bipartite graph that can be represented by an $m \times n$ coincidence matrix H with individual entries h_{ij}, where rows correspond to harvesters and columns correspond to servers. Viewing both harvesters and servers as resources that a spammer uses in order to send spam emails, a bicluster can be interpreted as the set of resources used by a spammer or group of spammers. To identify these biclusters, we perform spectral graph partitioning, also known as spectral clustering, on the bipartite graph of harvesters and spam servers. Since the graph is bipartite, the result is a set of biclusters containing both harvesters and servers. We refer to this method as spectral biclustering.

3.1 Spectral Biclustering

Spectral clustering makes use of the eigenvectors of a graph's adjacency matrix to partition the graph. We denote the adjacency matrix by the $v \times v$ matrix W with entries w_{ij} indicating the weight of the edge connecting vertices i and j. If there is no edge between vertices i and j, then $w_{ij} = 0$. Spectral clustering provides a relaxed solution to the following graph partitioning problem over X:

$$\text{maximize} \quad \frac{1}{k}\sum_{i=1}^{k} \frac{\mathbf{x_i}^T W \mathbf{x_i}}{\mathbf{x_i}^T D \mathbf{x_i}} \tag{1}$$

$$\text{subject to} \quad X \in \{0,1\}^{v \times k} \tag{2}$$

$$X 1_k = 1_v, \tag{3}$$

where k is the number of clusters to divide the graph into, $\mathbf{x_i}$ denotes the ith column of X, D is a diagonal matrix with entries $D(i,i) = \sum_j w_{ij}$, and 1_v denotes the all-one vector of length v. In short, the problem is one of finding an optimal graph partition which maximizes the ratio of the sum of edge weights between vertices in the same cluster C_i to the sum of edge weights between any two vertices where one vertex is in C_i.

Finding the optimal X is an NP-hard problem. The spectral clustering solution involves first relaxing constraint (2), allowing X to be continuous. The optimal continuous solution consists of the eigenvectors corresponding to the k largest eigenvalues of the matrix $\tilde{W} = D^{-1/2} W D^{-1/2}$. A near-optimal discrete solution is then obtained by thresholding these eigenvectors to satisfy (2) [11].

For a bipartite graph, the adjacency matrix has the special form

$$W = \begin{bmatrix} 0 & H \\ H^T & 0 \end{bmatrix} \tag{4}$$

where H is the $m \times n$ coincidence matrix of the bipartite graph. The eigenvalues and eigenvectors of W form a one-to-one correspondence with the singular values and singular vectors of H so we can work with H rather than W. Similarly, the eigenvalues and eigenvectors of $\tilde{W}$ form a one-to-one correspondence with the singular values and singular vectors of $\tilde{H} = D_1^{-1/2} H D_2^{-1/2}$, as noted in [3], where D_1 and D_2 are diagonal matrices such that $D_1(i,i) = \sum_j h_{ij}$ and $D_2(j,j) = \sum_i h_{ij}$. Let $U = [\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_k}]$ and $V = [\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_k}]$ denote the left and right singular vectors, respectively, corresponding to the k largest singular values of $\tilde{H}$. Let Z denote the $(m+n) \times k$ matrix given by

$$Z = \begin{bmatrix} D_1^{-1/2} U \\ D_2^{-1/2} V \end{bmatrix}. \tag{5}$$

The spectral biclustering algorithm can be implemented as follows. First, form the normalized coincidence matrix $\tilde{H}$. Next, compute the matrices of singular vectors U and V and form the matrix Z defined in (5). Finally, discretize Z to obtain a discrete partition matrix such that the (i,j)th entry is positive if vertex i is assigned to bicluster j and zero otherwise. The first m rows of Z correspond to the m harvesters, and the last n rows to the n servers. We refer interested readers to [11] for details on the discretization step.

3.2 Selecting the Number of Biclusters

Up to this point, we have assumed that k, the number of biclusters in which the data is divided, is known a priori. This is not the case, however. How to choose the optimal k is still an open problem, and many heuristics have been proposed. One that is particularly well-suited for spectral clustering is the eigengap heuristic, which suggests to choose k such that the largest eigenvalues of $\tilde{W}$, which correspond to the largest singular values $\sigma_1, \sigma_2, \ldots, \sigma_k$ of $\tilde{H}$, are very close to 1 but σ_{k+1} is relatively far from 1. A justification for this heuristic based on perturbation theory is given in [9]. We employ the eigengap heuristic for selecting k in this paper.

4. FINDINGS

We present our biclustering results for selected months. The results by month for the entire two-year period are available in [1].

4.1 Bicluster Sizes

The bipartite graph of harvesters and servers typically contains several extremely large biclusters, containing thousands to tens of thousands of vertices, forming the core of the graph. Recall that each bicluster is a set of harvesters and servers that are commonly used together. The other biclusters are usually much smaller and are usually only loosely connected to the core biclusters. This core-periphery structure has been observed in many real networks [5]. The distribution of the number of harvesters and servers in each bicluster over the two-year period is shown in Figure 3.

We believe the smaller biclusters correspond to individual spammers or groups of spammers working together. On the other hand, the larger biclusters are more difficult to interpret. They may contain resources that are heavily shared or publicly available, such as open relay servers. The existence of such vertices in the graph would create a blurring effect that negatively affects the ability to detect significant clusters in the graph using existing clustering methods such as spectral clustering [5]. Development of new approaches to detect significant clusters in the core of large graphs is an area for future work that shows promise for this application.

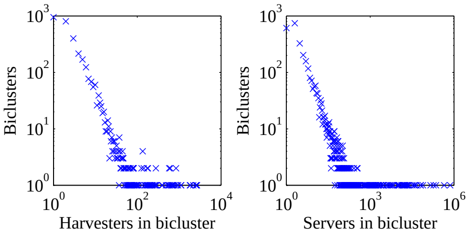

Figure 3: Bicluster size distribution for harvesters and servers.

4.2 Global Findings

One consistent observation from the biclustering results is that *biclusters appear to split into phishing and non-phishing biclusters*. That is, the biclusters tend to consist almost entirely of phishers or of non-phishers. This was found to be true for harvesters in [10]; in this study, we find it to be true for servers as well. Recall that the biclustering procedure was based entirely on harvester and server usage and was ignorant to the email contents. Specifically, the spectral biclustering algorithm was not provided with any information regarding phishing or non-phishing patterns.

The split can be seen in the bicluster interaction network shown in Figure 4, generated by force-directed layout in Cytoscape [8]. Each bicluster is represented by two vertices: a circular vertex corresponding to a cluster of harvesters and a triangular vertex corresponding to a cluster of servers. The size of a vertex denotes the number of harvesters or servers in that particular cluster, and the color of a vertex denotes the average phishing level of the harvesters or servers in the bicluster. Note that most biclusters are either black or white, showing that most biclusters do indeed consist almost entirely of phishers or of non-phishers. This suggests that most spammers are dedicated phishing or non-phishing spammers. We identify a bicluster as a phishing bicluster if it has more phishers than non-phishers and as a non-phishing bicluster otherwise.

Another interesting observation is that only the non-phishing biclusters display the core-periphery structure mentioned in Section 4.1, while the phishing biclusters tend to be more scattered. This is also visible in the bicluster interaction network shown in Figure 4. Since we believe the core biclusters correspond to heavily shared or publicly accessible resources, this suggests than many non-phishing spammers are using these shared or public resources while phishing spammers are not. Phishing biclusters are generally only weakly connected, suggesting that resource sharing between phishing spammers is minimal, possibly because phishing is a more dangerous activity than non-phishing from a legal standpoint.

We can obtain a quantitative representation of the split between phishing and non-phishing biclusters by looking at a contingency table comparing the number of phishers and non-phishers in phishing and non-phishing biclusters. The contingency tables for harvesters and servers in September 2006 are shown in Table 1. In this table we can observe the distribution of phishing and non-phishing harvesters and servers in phishing and non-phishing biclusters. There is clearly a large difference in the distribution of phishers and non-phishers in the two types of biclusters.

We find that most phishing and non-phishing harvesters are in phishing and non-phishing biclusters, respectively, over the entire two-year period. This also applies to the servers up to September 2006. Beginning in October 2006, many phishing servers are found

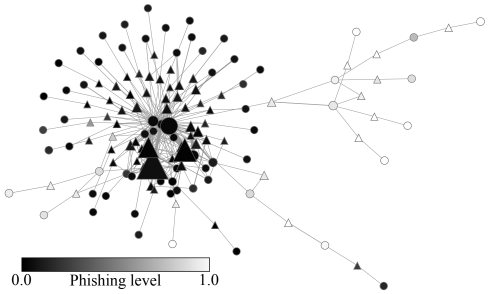

Figure 4: Bicluster interaction network for October 2006. Circles denote harvesters, and triangles denote servers.

	$P_{\text{harv.}}$	$N_{\text{harv.}}$
P_{cluster}	167	13
N_{cluster}	23	961

(a) Harvesters

	P_{server}	N_{server}
P_{cluster}	212	20
N_{cluster}	111	$37,553$

(b) Servers

Table 1: Contingency tables for September 2006. P and N denote phishing and non-phishing, respectively.

in non-phishing biclusters, as shown in Table 2. We find that these phishing servers belong to the extremely large biclusters, which are difficult to interpret as previously mentioned. These biclusters consist mostly of non-phishing servers and are identified as non-phishing biclusters as a result. Since servers are only connected to harvesters, this indicates that a few harvesters can be identified as mixed harvesters, which are associated with both phishing and non-phishing emails. The existence of these mixed harvesters suggests either that there are a few spammers who are sending both phishing and non-phishing spam or that a few phishing and non-phishing spammers are sharing lists of email addresses. Investigation of these mixed harvesters is an area for future work.

4.3 Local Findings

Examination of individual biclusters reveals several very interesting groups of spammers. In particular, we find that some biclusters contain harvesters or servers that have very similar IP addresses, suggesting that the resources may be physically close. Although this finding is not particularly surprising because large ISPs have IP ranges, it provides some validation for our biclustering results. One particular bicluster contains fifteen harvesters all with IP addresses under the 208.66.192/22 prefix owned by McColo Corp., a known rogue ISP that acted as a gateway to spammers until it was removed from the Internet in November 2008 [6]. Almost all of these fifteen harvesters are found in the same bicluster in each month. These harvesters were also found to be highly correlated in their temporal behavior in [10]. No similarity in IP addresses was found between the servers in this bicluster, indicating that the actual transmission of the spam emails was done in a distributed manner, possibly using botnets.

5. DISCUSSION

In this paper, we presented a biclustering analysis on the resource usage patterns of spammers. We identified biclusters as individual spammers or groups of spammers working together. We found that not many resources were used for both phishing and non-phishing, suggesting a split between phishing and non-phishing spammers. Additionally, the interaction network of phishing biclusters had a very different structure than that of non-phishing biclusters. We also discovered several biclusters consisting of resources with similar IP addresses, indicating physical proximity.

	$P_{\text{harv.}}$	$N_{\text{harv.}}$
P_{cluster}	155	4
N_{cluster}	37	$1,336$

(a) Harvesters

	P_{server}	N_{server}
P_{cluster}	230	23
N_{cluster}	$1,466$	$73,736$

(b) Servers

Table 2: Contingency tables for October 2006.

While our analysis has shown good preliminary results, it also leaves many questions to be answered. As discussed previously, there is the problem of interpreting the extremely large biclusters. Does a large bicluster actually correspond to a group of collaborating spammers or to many smaller groups that cannot be identified by existing clustering methods? Also, how can we interpret the large number of phishing servers being found in non-phishing biclusters beginning in October 2006? These are just a few of the questions that still require further investigation.

6. ACKNOWLEDGMENTS

We would like to thank Matthew Prince, Eric Langheinrich, and Lee Holloway of Unspam Technologies Inc. for providing us with the Project Honey Pot data. This work was partially supported by NSF grant CCF 0830490 and ONR grant N00014-08-1-1065.

7. REFERENCES

[1] Supplementary material. `http://tbayes.eecs.umich.edu/xukevin/spam_ceas2010/`.

[2] Project Honey Pot, 2010. `http://www.projecthoneypot.org`.

[3] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proc. ACM SIGKDD Conf. Knowl. Discov. and Data Mining*, 2001.

[4] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research*, 13(4):703–716, 2003.

[5] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proc. Int. World Wide Web Conf.*, 2008.

[6] J. Nazario. Third "bad ISP" disappears—McColo gone, 2008. `http://asert.arbornetworks.com/2008/11/third-bad-isp-dissolves-mccolo-gone/`.

[7] M. Prince, L. Holloway, E. Langheinrich, B. M. Dahl, and A. M. Keller. Understanding how spammers steal your e-mail address: An analysis of the first six months of data from Project Honey Pot. In *Proc. Conf. Email and Anti-Spam*, 2005.

[8] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:2498–2504, 2003.

[9] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[10] K. S. Xu, M. Kliger, Y. Chen, P. J. Woolf, and A. O. Hero III. Revealing social networks of spammers through spectral clustering. In *Proc. IEEE Int. Conf. Commun.*, 2009.

[11] S. X. Yu and J. Shi. Multiclass spectral clustering. In *Proc. IEEE Int. Conf. Comput. Vision*, 2003.