InferIP: Extracting actionable information from security discussion forums

Joobin Gharibshah*, Tai Ching Li*, Maria Solanas Vanrell*, Andre Castro*,
Konstantinos Pelechrinis†, Evangelos E. Papalexakis* and Michalis Faloutsos*
* University of California - Riverside, CA
Email: {jghar002,tli010,msola004,acast050,epapalex,michalis}@cs.ucr.edu
† School of Information Sciences, University of Pittsburgh, Pittsburgh, PA
Email: kpele@pitt.edu

***Abstract*—How much useful information can we extract from security forums? Many security initiatives and commercial entities are harnessing the readily public information, but they seem to focus on structured sources of information. Our goal here is to extract information from hacker forums, whose information is provided in ad hoc and unstructured ways. Here, we focus on the problem of identifying malicious IPs addresses, when these are being reported in the forums. We develop a method to automate the identification of malicious IPs with the design goal of being independent of external sources. A key novelty is that we use a matrix decomposition method to extract latent features of the behavioral information of the users, which we combine with textual information from the related posts. As key design feature, our technique can be applied to different language forums since it relies on a simple NLP solution in combination with behavioral features. In particular, our solution only needs a small number of keywords in the new language plus the user's behavior captured by specific features. We also develop a tool to automate the data collection from security forums. We collect approximately 600K posts from 3 different forums. Our method exhibits high classification accuracy, while the precision of identifying malicious IP in post is greater than 88% in all three sites. Furthermore, by applying our method, we find up to 3 times more potentially malicious IPs than compared to the reference blacklist VirusTotal. As the cyber-wars are becoming more intense, having early accesses to useful information becomes more imperative to remove the hackers first-move advantage, and our work is a solid step towards this direction.**

Keywords: Security, Online communities mining

I. Introduction

How can we remove the advantage of surprise from malicious hackers? This is the overarching goal of this project. In this work, we address a specific question. In particular, we want to extract as much useful information from hacker/security forums as possible in order to perform (possibly early) detection of malicious IPs, e.g., prior to their appearance on blacklists. The latter can exhibit large delays in their update and hence, new ways for labeling malicious IPs are needed [8]. In this study we will use the term "hacker forums" to describe online forums with a focus on security and system administration. Interestingly, we can classify these forums into categories: (a) main stream forums, like `wilderssecurity`, and (b) "fringe" forums, like `offensivecommunity`, where we find users with names like *satan911*. Some of the fringe forums have been known to have hackers boast of attacks they have mounted, or sell tools and infrastructure for malicious purposes (think rent-a-botnet). For example, in our dataset there is a post that mentions "I give you a second server to have your fun with. Multiple websites on this server. So let's see if anyone can actually bring down the server". Right after that the hacker posted the IP, username and password for anyone to access the server. In fact, there is a *show-off* section in these forums for people to broadcast their hacking "skills".

To reiterate, the central theme in our work is to develop techniques for extracting information from a security forum with the goal of informing a security analyst. The particular problem of our study is to identify malicious entities, and more specifically malicious IPs. Formally, our problem is as follows:

Key Question: Malicious IP Detection. Given a set of posts $\mathcal{P}_F$ that may contain IP addresses and users $\mathcal{U}_F$ of a security forum F, as well as, the features Φ_p, $\forall p \in \mathcal{P}_F$ and Φ_u, $\forall u \in \mathcal{U}_F$ for the posts and the users respectively, can we determine if a given IP address i is malicious or not?

The set of features $\mathcal{P}_F$ includes attributes such as the text of the post, the posting user, the time of post, etc., while $\mathcal{U}_F$ includes information such as the date of a user joining the forum, the number of posts the user has made etc.

TABLE I: Extracting useful information; Number of malicious IPs found by InferIP and not by VirusTotal.

		IP found by	
Dataset	Total IP	Virus Total	InferIP only
Wilders Security	4338	216	**670**
Offensive Community	7850	339	**617**
Ashiyane	8121	133	**806**

Most previous studies in this area have focused on mining structured information sources, such as security reports. In fact many efforts focus on addressing security problems using knowledge obtained from the web, as well as, social and information networks, these efforts are mainly focused on analyzing structured sources (e.g., [9]). However, studies assessing the usefulness of (unstructured) information in online forums have only recently appeared (e.g., [14]). These studies are mostly exploratory in that they provide evidence of the usefulness of the data in the forums, but do not provide a systematic methodology or ready-to-use tools, which is the goal of our work. We discuss existing literature in more detail later in this section.

The motivation of our work is to enhance our security knowledge and to complement, and not to replace, existing efforts for detecting malicious IPs. For instance, IP blacklists enlist an IP as malicious after a number of reports above a pre-defined threshold have been made for the specific address. Depending on the threshold and the reactivity of the affected users/systems, this might take several days, weeks or months to happen. Therefore, a system whose core is the solution of our Key Question can identify and recommend (potentially) malicious IP address to blacklist services and firewalls.

We propose , a systematic method to identify malicious IPs among the IP addresses which are mentioned in security forums. A key novelty is that we use the behavioral information of the users, in addition to the textual information from the related posts. We customize and use a Sparse Matrix Regression method on this expanded set of features. By design, our framework is applicable to forums in different languages as it relies on and the behavioral patterns and keywords and not a complex language-specific NLP technique. From a technical point of view the challenge in designing a solution to our Key Question is most IPs mentioned in these forums are not malicious. We show that our system can add a significant number of previously unreported IP address to existing blacklist services.

We develop a customizable and flexible crawler for forums, that only requires a simple specification file. Using our crawler, we collect data from three forums, two English and one in Farsi for a total number of more than 30K users and 600K posts. We use VirusTotal [3] as our reference blacklist IP addresses, since it is an aggregator, and combines the information from over 60 other blacklists and resources. Our results can be summarized into the following points:

a. Our method exhibits precision and recall greater than 88% and 85% respectively, and an accuracy over *malicious* class above 86% in the 10-fold cross validation tests we conducted for the three different forums. In partially answering our Key Question, if our method labels a currently non-blacklisted IP as malicious, there is a high chance that it is malicious, given our high precision.

b. Our method identifies three times more malicious IPs compared to VirusTotal. We find more than 2000 potential malicious IPs that were never reported by VirusTotal among our three forums.

II. DATA COLLECTION AND BASIC PROPERTIES

We have collected data from three different forums relevant to our study; (i) WildersSecutiry [4], (ii) OffensiveCommunity [2], (iii) Ashiyane [1]. The first two forums are mainly written in English, while the last forum is an Iranian forum, in Farsi[1]. Some basic statistics for these forums are presented in Table II. OffensiveCommunity and Ashiyane are two fringe forums in different languages. In these forums there is a section where people openly boast about their achievement in hacking. They share their ideas and *tutorials* on how to break into vulnerable networks. On the other hand, WildersSecurity as a main stream forum is mostly used to protect non-experts against attacks such as browser hijacking, and provide solutions for their security problems. For completeness, we present some of the terms we use here. A user is defined by a login name registered with the site. The term post refers to a single unit of content generated by a user. A thread refers to a collection of posts that are replies to a given initiating post.

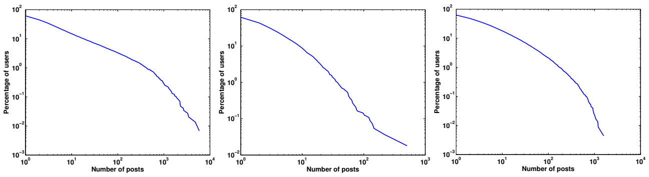
TABLE II: The collected forums.

Forum	Threads	posts	users	Active days
Wilders Security	28661	302710	14836	5227
Offensive Comm.	3542	25538	5549	1508
Ashiyane	67004	279309	22698	4978

Figures 1 and 2 present the cumulative complementary distribution function for the number of posts per user and the number of threads per users respectively. As we can see in all the cases the distributions are skewed, that is, most of the users contribute few posts in the forums and engage with few threads. In Wilders Security 85% of users post less than 10 posts each, while 5.2% of the users post more than 50 posts. 70% of the users post in only one thread and only 8% of the users are active in more than 10 threads This skewed behavior is typical for online users and communities [7]. We develop features to capture aspects of both these user properties, as we will see next. Due to space limitations, we cannot present plots for more features that we use in our classification.

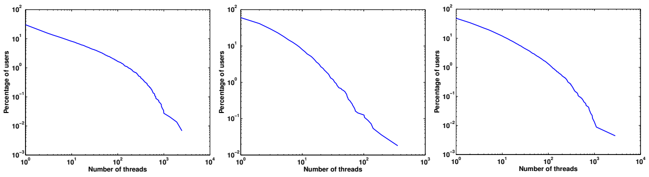
Groundtruth for training and testing. In order to build and evaluate, our model we need to obtain a reasonably labeled dataset from IP addresses that appear in the posts of the security forums. For that, we use the VirusTotal service and assign malicious labels to an IP that has been reported by this service. The number of malicious IPs that we have use with the corresponding posts are shown in table I as the IP found by VirusTotal. Note that the absence of a report on VirusTotal does not necessarily mean that the IP is benign. However, a listed IP address is most likely malicious, since VirusTotal as most blacklist sites require a high threshold of confidence for blacklisting an address. This way, we find in total 688 malicious IPs for our forums as shown in Table I.

Using this labeling process we have collected all the IPs that have appeared on our forums prior to their report on VirusTotal. For building our model, we also randomly select an equal number of IPs that have not been reported as malicious and via manual inspection further assess their status. Finally, for every security forum we have a different dataset and hence, we build a different model.



(a) WildersSec. (b) OffensiveCo. (c) Ashiyane

Fig. 1: CCDF of the number of posts per user (log-log scale).



(a) WildersSec. (b) OffensiveCo. (c) Ashiyane

Fig. 2: CCDF of the number of thread per user (log-log scale).

III. INFERIP: MALICIOUS IP DETECTION

We propose a method to identify whether an IP address within a post is malicious. For example, although many users report a malicious IP address, such as one that is attacking the

[1] Our software and datasets will be made available at: https://github.com/hackerchater/

user's network, there are also users that will mention a benign IP address when people discuss about network tutorials like setting up *Putty* or initiating a *SSH* connection.

While this task is simple for a human, it is non-trivial to automate. Adding to the challenge, different communities use different terminology and even different languages altogether (english and farsi in our case). In order to overcome these challenges, we use a diverse set of features and build a model to identify IPs that are potentially malicious.

Our approach consists of four steps that each hides non-trivial novelties:

Step 1: We consider the user behavior and extract features that profile users that post IP-reporting posts.

Step 2: We extract keywords from the posts and use information gain to identify the 100 most informative features.

Step 3: We identify meaningful *latent feature sets* using an unsupervised co-clustering approach [12].

Step 4: We train a classifier using these *latent feature sets* using 10-fold cross validation.

We describe each step in more detail.

Step 1: Behavioral Features. We associate each user of the forum with a set of 11 features that capture their behavior. In particular:

- Number of posts; the total number of posts made by the user
- Number of threads; the total number of threads the user has contributed to
- Number of threads initiated; the total number of threads initiated by the user
- Average thread entropy; the average entropy of the user distribution of the threads in which the user has contributed to
- Number of active days; the number of days that the user generates at least one post
- Average day entropy; the average entropy of the user distribution of the posts made on the days that the user is active
- Active lifetime; the number of days between the first and the last post of the user
- Wait time; the number of days passed between the day the user joined the forum and the day the user contributed their first post
- Average post length; the average number of characters in the user's posts
- Median post length; the median number of characters in the user's posts
- Maximum post length; the number of character's in the user's longest post

Step 2: Contextual Features. Apart from the aforementioned behavioral features we also include features related with the context in which an IP address appears within a post. In particular, we consider the frequency of the words (except stop-words) in the posts. Words that are frequent only in few documents (posts in our case) are more informative than those that appear frequently on a larger corpus [13]. To this end, we use TF-IDF to weight the various words/terms that appear in our data. After calculating the frequency and the corresponding weights of each word in the dataset we end up with more than 10,000 features/terms. Hence, in the next step we select discriminative features by extracting latent features.

TABLE III: Selecting a classifier: overall accuracy.

Forum	Naive Bayes	3NN	Logistic regression
Wilders Security	91.9%	87.1 %	94.8%
Offensive Comm.	84.1%	83.2%	86.5%
Ashiyane	85.1%	82.3%	94%

TABLE IV: InferIP evaluation: 10-fold cross validation evaluation (using Logistic Regression).

Forum	Instances	Precision	Recall	ROC Area
Wilders Security	362	0.9	0.94	0.96
Offensive Comm.	342	0.88	0.85	0.91
Ashiyane	446	0.9	0.92	0.92

We begin by performing feature selection in order to identify the most informative features by applying the information gain framework [15]. Furthermore, in order to avoid overfitting we pick a random subset of posts from the whole dataset and select the highest ranked features based on *Information Gain* score. In this way, a subset of discriminative keywords, 100 in our model, are selected. It turns out that each user uses only a small number of those words, resulting in a sparse dataset which we wish to exploit in our model.

Step 3: Identifying latent feature sets. We also like to leverage latent similarities of different posts in some of the dimensions spanned by post features and behavioral features for the writer of the post. Essentially, we seek to identify groups of highly similar posts under a small number of features, which does not necessarily span the full set of features. The reason why we wish to pinpoint a subset of the features instead of the entire set is because this way we are able to detect subtle patterns that may go undetected if we require post similarity across all the features. We call those sets of feastures *latent feature sets* . To this end, we apply a soft co-clustering method, Sparse Matrix Regression (SMR) [12], to exploit the sparsity and extract latent features of the post containing IPs. Given a matrix $\mathbf{X}$ of posts $\times$ features, its soft co-clustering via SMR can be posed as the following optimization problem:

$$\min_{\mathbf{a}_r \geq 0, \mathbf{b}_r \geq 0} \|\mathbf{X} - \sum_{r}^{R} \mathbf{a}_r \mathbf{b}_r^T\|_F^2 + \lambda \sum_{i,r} |\mathbf{a}_r(i)| + \lambda \sum_{j,r} |\mathbf{b}_r(j)|$$

where $\mathbf{a}_r$ and $\mathbf{b}_r$ are vectors that "describe" co-cluster r, which we explain below. Each $\mathbf{a}_r$ is a vector with as many dimensions as posts. Each value $\mathbf{a}_r(i)$ expresses whether post i is affiliated with co-cluster r. Similarly, $\mathbf{b}_r$ is a vector with as many dimensions as features, and $\mathbf{b}_r(j)$ expresses whether feature j is affiliated with with co-cluster r. Parameter λ controls how sparse the co-cluster assignments are, effectively controlling the co-cluster size. As we increase λ we get sparser results, hence cleaner co-clustering assignments. We tune λ via trial-and-error so that we obtain clean but non-empty co-clusters, and we select $\lambda = 0.01$ in our case.

Step 4: Training the model. We subsequently train a number of classifiers using the selected features based on a matrix. In particular, we examine (a) a Naive Bayes classifier, (b) a K-Nearest Neighbor classifier and (c) a logistic regression classifier. Our 10-fold cross validation indicates that the Logistic regression classifier outperforms kNN and Naive Bayse, achieving high accuracy, precision and recall (see Table III).

Applying InferIP on the forums. Having confidence in our classifier, we want to apply it on the posts of the forums except the ones that we used in our groundtruth. Naturally, we use the logistic regression classifier as it exhibits the best performance. With InferIP, we find an additional 670 malicious IPs in `WildersSecurity`, and 617 in `OffensiveCommunity` 806 in `Ashiyane` (see Table I). In other words, InferIP enables us to find three times additional malicious IPs in total compared to the IPs found on VirusTotal. It is interesting to observe that this factor varies among our three sites. For Ashiyane, our method finds roughly 6 times additional malicious IPs. With a precision of roughly around 90% and considering small amount of *False Positive* rate, our method can add a significant number of malicious IPs to a blacklist. Using the limited manual inspection, we confirm that the precision of the method on out of sample data is in the order of 88%.

IV. Related Work

We briefly discuss two categories of relevant research.

a. Analyzing structured security sources. There is a long line of research studying the ontology of cyber security and the automatic extraction of information from structured security documents (e.g., [9], [6]). This work is complementary to ours as it focuses on different information sources with different challenges.

b. Analyzing online security forums. Recently security forums have been the focus of various studies that showcase the usefulness of the information present in security forums. For example, Motoyama *et al.* [11] present a comprehensive statistical analysis in underground forums. Others studies focus on the users' classification or the discovery of the relationships between the forum's members [16], [5]. Extracting different discussion topic in the forums and classifying the language of the codes posted in the forum has been done in [14]. Contrary to these studies, our work emphasizes on the development of automated systems that actually exploit the wealth of information in these security forums in order to enhance security. Similar to detecting malicious users on commenting platforms has been done on [10].

V. Conclusion

The overarching take away message from our work is that there could be a wealth of useful information in security forums. The challenge is that the information is unstructured and we need novel methods to extract that information. A key insight of our work is that using behavioral and text-based features can provide promising results.

In support of this assertion, we develop a systematic method to extract malicious IP addresses from chatter in security forums. We utilize both behavioral as well as textual features and show that we can detect malicious IPs with high accuracy, precision and recall using simple classifiers. Our results at Table I are promising. We find three times as many additional malicious IPs as the original malicious IPs identified by `VirusTotal`. While this does not mean that all of the IPs that we find are malicious, our high precision (hovering around 90% in Table IV) suggests that most of them are indeed malicious.

In the future, we plan on extending our work to enhance other security tasks by extracting as much useful information as possible from security forums. Our first goal is to detect malicious URLs mentioned in the forums. Our second and more ambitious goal is to identify the emergence of new malware, threats, and possibly attacks, which we expect to see as large numbers of panicky posts. Finally, our goal is to identify malicious users, since interestingly, some users seem to be promoting or maybe even selling hacking tools.

VI. Acknowledgments

The authors thank the anonymous reviewers for their useful comments. This material is based upon work supported by the Bourns College of Engineering at University of California, Riverside, and DHS ST Cyber Security (DDoSD) HSHQDC-14-R-B00017 grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or other funding parties.

References

[1] Ashiyane. http://www.ashiyane.org/forums/.

[2] Offensive community. http://www.offensivecommunity.net.

[3] Virustotal. http://www.virustotal.com.

[4] Wilders security. http://www.wilderssecurity.com.

[5] A. Abbasi, W. Li, V. Benjamin, S. Hu, and H. Chen. Descriptive analytics: Examining expert hackers in web forums. In *2014 IEEE Joint Intelligence and Security Informatics Conference*, pages 56–63, Sept 2014.

[6] C. Blanco, J. Lasheras, R. Valencia-García, E. Fernández-Medina, A. Toval, and M. Piattini. A systematic review and comparison of security ontologies. In *2008 Third International Conference on Availability, Reliability and Security*, pages 813–820, March 2008.

[7] P. Devineni, D. Koutra, M. Faloutsos, and C. Faloutsos. If walls could talk: Patterns and anomalies in facebook wallposts. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 367–374, New York, NY, USA, 2015. ACM.

[8] H. Hang, A. Bashir, M. Faloutsos, C. Faloutsos, and T. Dumitras. "Infect-me-not": A user-centric and site-centric study of web-based malware. In *IFIP Networking*, pages 234–242, May 2016.

[9] M. Iannacone, S. Bohn, G. Nakamura, J. Gerth, K. Huffer, R. Bridges, E. Ferragut, and J. Goodall. Developing an ontology for cyber security knowledge graphs. In *Proceedings of the 10th Annual Cyber and Information Security Research Conference*, CISR '15, pages 12:1–12:4, New York, NY, USA, 2015. ACM.

[10] T. C. Li, J. Gharibshah, E. E. Papalexakis, and M. Faloutsos. Trollspot: Detecting misbehavior in commenting platforms. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, 2017.

[11] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker. An analysis of underground forums. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11, pages 71–80, New York, NY, USA, 2011. ACM.

[12] E. E. Papalexakis, N. D. Sidiropoulos, and R. Bro. From k-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors. *IEEE transactions on signal processing*, 61(2):493–506, 2013.

[13] J. Ramos. Using TF-IDF to determine word relevance in document queries. In *Instructional Conference on Machine Learning*, 2003.

[14] S. Samtani, R. Chinn, and H. Chen. Exploring hacker assets in underground forums. In *IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 31–36, May 2015.

[15] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[16] X. Zhang, A. Tsang, W. T. Yue, and M. Chau. The classification of hackers by knowledge exchange behaviors. *Information Systems Frontiers*, 17(6):1239–1251, Dec. 2015.