URLytics: Profiling Forum Users from their Posted URLs

Ben Treves
UC Riverside
btrev003@ucr.edu

Md Rayhanul Masud
UC Riverside
mmasu012@ucr.edu

Michalis Faloutsos
UC Riverside
michalis@cs.ucr.edu

***Abstract*—Online forums contain a substantial amount of data, but very few studies have focused on mining the URLs posted by users. How can we fully leverage these posted URLs to extract as much information as possible about forum users? We perform a systematic study for extracting as much information as possible about forum users via their URL posting behavior. Within this study we develop a series of tools to analyze the data. Given a forum, we extract the following information: (a) basic statistics and a profile of the forum, (b) a profile for each user based on their referral to accounts in other platforms, (c) identification of communities within the forum, and (d) detection of malicious behavior. Most prior works focus on analyzing the text found in user posts rather than on URLs themselves, as we do here. In our study, we analyze three online security forums and find interesting results: (a) we identify 7% of the users posting social media links on other platforms, (b) we detect 148 groups of users that engage in communities on external social media platforms, (c) we expose 139 malicious users that collectively posted 328 malicious URLs. Additionally, we identify 17 groups with membership spanning across multiple forums, and discover numerous other groups that engage in coordinated malicious behavior. Our work is a significant step towards an all-encompassing system for profiling forum users at large.**

***Index Terms*—Online Forums, Social Media, User Profiling, Group Clustering, Malicious Behavior Detection.**

Introduction

Online forums hide a wealth of information that has not been studied as much as other social media until recently. We use the term **forum** to describe thread-driven online discussion platforms. There are 100K [1] online forums on the internet, with some having over 500M monthly active users [2]. This unprecedented amount of user generated content contains a wealth of information. URLs that users post can help us connect information from other sources such as social media, educational resources, news outlets, etc. with the activity of the users and even the forum itself.

The problem we address here focuses on the information that we can extract from the URLs found in forum posts. By contrast, most prior works focus on analyzing the text in the posts. The challenge is to identify the type and amount of information we can extract from URLs alone. Here we take a user-centric approach. Specifically, given one or more online forums, the desired output is as much information as we can extract for each user. This includes information about who they are and whether they engage in malicious activities. In our study, we define malicious behavior as sharing URLs in that enable phishing, spamming, or malware infections.

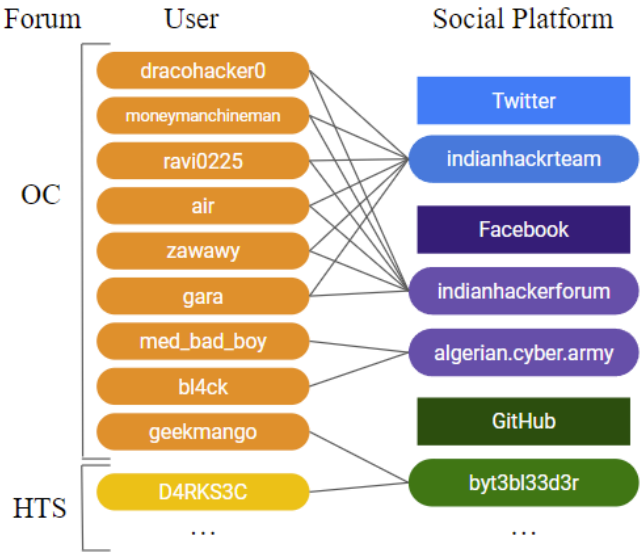

Figure 1: Group affiliation of users across forums (left) and social platforms (right). Groups can form across forum boundaries and reach cross-platform social communities.

There are many types of analysis we can perform, but to maintain a focused scope and reasonable paper length, we focus on what we can discover about interconnections between forums and social media, as we explain in more detail later in the paper. In our analysis, we focus on the URLs themselves that users post. We examine some post content around the URL for identity disambiguation purposes, but otherwise avoid NLP techniques altogether.

Extracting information out of URLs alone is a non-trivial task. The URLs often lead to broken sites due to the age of the data, which means we could not generally open a URL and crawl the linked web page. Additionally, URL extraction from post data is not perfect, as sometimes URLs have typos and sometimes typos are flagged as URLs. Therefore, there must be a balanced trade-off between the precision and the recall of the URL extraction method.

As our key contribution, we conduct a systematic study, *URLytics*, to highlight the wealth of information hiding in URLs posted in forums. We develop a suite of techniques that streamline the analysis of URLs. We show that posted URLs can help us establish connections between different social media and online platforms. Specifically, we analyze

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 IEEE

Table I: Dataset Overview

Forum Name	Total Users	Total Posts	Total URLs	URL Users	Social URLs
OC	5499	25538	22722	821	269
HTS	9423	68464	13880	2264	727
EH	2970	50908	5544	636	145
WLD	14660	302711	7439	1031	246
Total	32552	447621	49585	4752	1141

URLs to: (a) perform identity disambiguation by connecting users across different platforms, (b) identify groups and clusters of cross-platform users, and (c) detect malicious behaviors, including phishing, spamming, and spreading malware. For our identity disambiguation, we develop an automated approach that combines: (a) string matching using Levenshtein and Jaro distance and (b) context awareness from the post of the URL. Our work is among the first systematic studies that focus on mining URLs posted in forums.

We apply our approach on data from four security forums. The data we use spans from 2002 to 2022 and contains 33K users and 450K posts. We opt to analyze this niche community of interest to security analysts since these forums often include malicious hackers. The key results of our work can be summarized in the following points.

A. Social media URL posting statistics. We find that different forums vary significantly in their social media URL posting profile. Looking at the top 3 most prevalent social media platforms in our data, OC is dominated by Facebook activity, HTS is split equally between GitHub and Twitter usage, while in EH Twitter is much more prevalent. See Figure 2 for a visualization of these distributions.

B. Identification of users across platforms. We are able to establish connections between forum users and their linked social media accounts. We find that 7% of the social media accounts linked by forum users belong to the users posting them with high confidence.

C. Community discovery. We find 148 groups of users that share the same social media URL links. 17 of these groups include members from multiple forums. Manual inspection of the most populated of these groups revealed groups engaging in activities that range from educational programming competitions to political hacking efforts.

D. Malicious activities. We find 139 malicious users that collectively share 328 malicious URLs. These are URLs that would have infected a user that clicked on them, and range in severity from phishing to spreading malware. Later in the paper we explain how users can share unclickable malicious links in a benign way. We ignore these unclickable links for our study.

Dataset and Terminology

In our study, we focus on three online security forums: Offensive Community [3], Ethical Hacker [4], and Hack This Site [5], as well as a fourth online security forum to verify our results, Wilders Security [6]. These forums contain posts from 2002 to 2022. A given post consists of the text that a user writes in a single post on a single forum as well as the metadata of that post (user ID, post date, user title, etc.) A summary of our forum dataset can be found in Table I.

For the rest of the paper, we refer to the forums Offensive Community, Ethical Hacker, Hack This Site, and Wilders Security as OC, EH, HTS, and WLD, respectively.

Methodology

In this section we describe our methods for extracting the following information from user URL posting behavior: (a) user identification across platforms, (b) group affiliation, and (c) malicious behavior. We mainly focus on Facebook, Twitter, GitHub, and YouTube for our social media analysis as other social media platforms were not nearly as prevalent in our extracted forum data. We identify social media URLs by detecting if they link to the top social media platforms, ranked by popularity [7].

A. User identification across platforms. When we encounter a user that posted a social media URL, we want to find out if the linked social media account belongs to the user that posted it. For this purpose, we use a combination of two methods: (i) string matching usernames, (ii) context around the link.

(i) String matching usernames. If a forum user and a social media account that they linked have similar usernames, it is very likely that they are the same user. First, we attempt to extract the username from the posted social media URL. Here, we do not use YouTube URLs because they do not contain the account username as part of the URL. For string matching, we use two string edit-distance algorithms, Levenshtein and Jaro distance, with thresholds of 4 and 0.75, respectively. We chose these thresholds as a result of our manual test of a random sample of 50 username pairs from our data, in which each algorithm yielded at most 2 erroneous matches or mismatches with these values. If one of the algorithms yields a match, we conclude that the forum user is connected to the linked social media account.

(ii) Context around the link. If the string matching method does not yield a match, we check for possessive words (such as ”my”, ”mine”, or ”our”) in the text around the URL. The existence of such words in short vicinity of the URL implies that the user is claiming ownership of the URL. We manually tested a randomly selected sample of 15 users from each forum and found that a maximum word distance of 20 words from the URL is optimal for associating a social media URL with a user. Essentially, if a user writes a possessive word within 20 words of a social media URL, we associate that user with the linked social media account.

B. Group affiliation. One of the key tasks of our work is to find groups that users are affiliated with solely based on their URL posting behavior. For the purpose of discovering these groups, we propose a bipartite graph structure where the left-hand side consists of forum users and the right-hand side is social media accounts. Whenever a user posts a social media URL, they are connected to that social media account in our graph. This allows us to formulate groups of forum

users that might not be interacting directly, but are at least engaging in the same communities off-platform. We split the forum side of the graph into three sections for each of our three forums, and the right side of the graph into three sections for each of our targeted social media platforms. A snippet showing a few groups from our graph can be seen in Figure 1.

C. Malicious behavior. We identify malicious users and the type of malicious behavior they engage in using their URL posting behavior. To detect malicious URLs, we utilize an ensemble of 7 well known malware detection engines (including Webroot and Avira) to analyze each URL in our dataset [8]. When a detection engine flags a URL as malicious, it returns the type of malicious behavior that URL exhibits (such as Malware or Phishing), which we use to flag the user that posted that URL with.

We only focus on URLs that users intend others to click on. It is a commonly accepted practice to mask malicious URLs as non-clickable when they are posted for benign purposes. Replacing the protocol *http* or *https* with *hxxp* or *hxxps*, respectively, is widely used by the security community to obfuscate URLs [9]. Supporting this claim, we find 84 URLs in our data that are masked as *hxxp*. Therefore, we conclude that when users post clickable, malicious URLs they are in fact engaging in malicious malicious activity, and are flagged by our detection script appropriately.

RESULTS

The main contribution of this paper is to show that there is a wealth of information in URLs that helps us establish connections between different platforms and provides interesting information. In this section, we report our findings of the following information: (a) social media URL posting statistics, (b) identification of users across platforms, (c) community discovery, and (d) malicious activities.

A. Social media URL posting statistics. We find non-trivial connections between online forums and external social media platforms, the intensity and diversity of which varies by forum. As can be seen in Figure 2, each forum tends to favor a different distribution of social media URL sharing. We discover that OC users favor Facebook URLs, EH users favor Twitter URLs, and HTS users post almost equal GitHub and Facebook URLs, with rarely any Twitter presence.

B. Identification of users across platforms. We are able to establish connections between forum users and the linked social media accounts. For this purpose we extract the username from the URL shared by the user and perform various string matching analyses to determine if the user is sharing their own account. We find that 7% of the social media accounts linked by forum users belong to the users posting them. We verify our result using two different string matching algorithms and a post context method checking for user possession of the URL, as described in the Methodology section.

C. Community discovery. We identify communities of users across different forums by analyzing their member

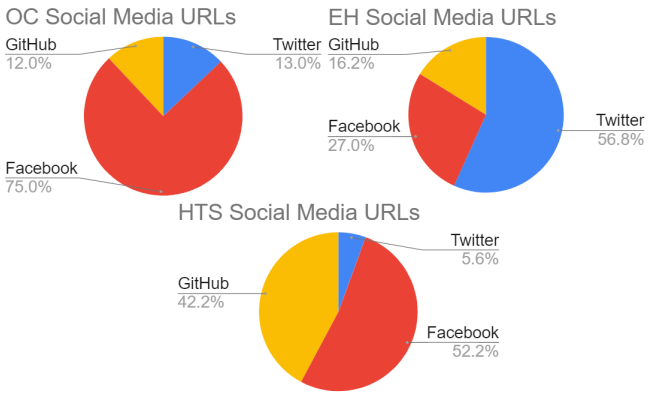

Figure 2: Distribution of social media URLs across our security forums, OC, EH, and HTS. Facebook shown in red, Twitter shown in blue, and GitHub shown in yellow. Note that each forum has unique tendencies to use specific social media platforms, such as EH favoring Twitter while OC heavily favors Facebook.

users' social media links. We discuss the key results from our community discovery efforts in the following paragraphs. A snippet of our results can be seen in Figure 1.

(i) *Member discovery.* We discover the members of external communities operating on social platforms such as Facebook and GitHub. We consider only those external communities that have at least two members from our forums, otherwise the linked URL is potentially a single account. Using this definition, we find 148 potential communities of users across the forums in our dataset.

(ii) *Cross-platform communities.* While performing community member discovery, we captured multiple groups that span across forums and even more groups with overlapping member bases. Out of the 148 potential communities, we detect 17 groups that consist of users from different forums. Next we discuss standout results from these findings.

(iii) *Peculiar and unique groups.* We manually inspect the most populated communities and uncover interesting standout results. We find members of a group across platforms known as *byt3bl33d3r* on GitHub, which has over 5K followers and includes hundreds of code contributors. It hosts a variety of tools and scripts designed for exploitation, mainly around networking, making the forum users that are part of this group potentially malicious. We also find multiple politically motivated hacking groups, such as *IndianHackrTeam* and *algerian.cyber.army*, each having over 1K members on Facebook, with *IndianHackrTeam* also having mirror communities on other platforms. Figure 1 shows the cross-platform nature of these groups. Additionally, we also discover many benign groups, such as a community on GitHub centered around programming competitions, that spans multiple users across the forums in our dataset.

D. Malicious activities. Using our ensemble of malware detection engines, described in the Methodology section, we discover 139 malicious users that collectively posted 328 malicious URLs across the forums in our dataset. These users engaged in the sharing of malicious links that range in severity from phishing to spreading infectious malware. Of the 139 detected malicious users, 45 of them are flagged

for spreading malware. Investigating this list of malware-spreading users further, we discover several high profile users, including user *aabee* that, in addition to posting two links to malware, illegally leaked several hundreds of online accounts and passwords.

Discussion

Combining URLs with text analysis. Here we focus on URLs alone in order to highlight the information that they contain. In the future, we intend to study URLs and the text of the post together. We expect that the combined analysis could provide even more extensive and detailed information. Note that here we only used the post text in direct relation to the URL.

YouTube and broken links. When performing our user identification analysis, we do not consider YouTube URLs. We find it difficult to extract useful username information from them in the same method we use with other social media links. This is due to YouTube links not always showing the account username in the URL itself. Additionally, most of the YouTube links led to invalid destinations, suggesting that the linked videos or accounts are are suspicious. We plan to investigate this lead in future work.

Related Work

Profiling users of online forums via their URL posting behavior is quite a niche field that received relatively little attention from researchers at large. Most of the existing studies can be grouped into three separate categories.

A. User profiling studies. The studies in this category profile users of a forum based on the post content and demographic attributes provided by the users. One such study finds the political preferences of twitter users by analyzing the textual content of their tweets [10]. Other studies link accounts across platforms by analyzing user writing styles [11] and exploring user generated metadata [12] to create connections.

B. Forum activity detection studies. The studies in this category focus on extracting activities from discussion forums by analyzing the posts generated by forum users. For identifying malicious activities, previous work has utilized analysis of textual data using machine learning or other NLP techniques [13], [14], [15]. Other efforts utilize metadata found in posts to identify interesting events throughout a forum's lifespan [16].

C. URL analysis studies. One URL analysis study classifies URLs as malicious or benign with the help of Naive Bayes and Support Vector model [17]. A more recent work uses URL specific features such as bag of words in a URL and the URL length and web page features to detect phishing URLs [18].

Conclusion

As our key contribution, we conduct a systematic study, *URLytics*, to highlight the wealth of information hiding in URLs posted in forums. Using an ensemble of string matching algorithms and the context around social media URLs we can reliably match usernames extracted from social media URLs to the users posting them on forums to identify users across platforms. By grouping forum users together by the social media accounts that they link to, we are able to identify 148 groups of users on our forums, 17 of which have cross-forum membership. These communities range from malicious political hacking groups to benign programming competition groups. We discover 139 malicious users that posted 328 URLs that are flagged as malicious by our ensemble of credible malware detection engines. Our work is a significant step towards an all-encompassing system for profiling forum users at large.

Acknowledgements

This work was supported by NSF SaTC Grant No. 2132642.

References

[1] Forum software usage distribution on the entire internet. [Online]. Available: https://trends.builtwith.com/cms/forum-software/traffic/Entire-Internet
[2] Best and most popular forums message boards online communities. [Online]. Available: https://it-maniacs.com/best-and-most-popular-forums-message-boards-and-online-communities-top-30/
[3] Offensive community. [Online]. Available: http://offensivecommunity.net/
[4] Ethical hacker. [Online]. Available: https://www.ethicalhacker.net/
[5] Hack this site. [Online]. Available: https://www.hackthissite.org/
[6] Wilder security. [Online]. Available: http://www.wilderssecurity.com/
[7] Global social media statistics. [Online]. Available: https://datareportal.com/social-media-users
[8] Advanced threat prevention and detection. [Online]. Available: https://metadefender.opswat.com/
[9] H. Salgado, "The 'hxxp' and 'hxxps' uri schemes," *Internet Engineering Task Force*, 2017.
[10] A. Boutet, H. Kim, and E. Yoneki, "What's in your tweets? i know who you supported in the uk 2010 general election," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 6, no. 1, 2012, pp. 411–414.
[11] T. N. Ho and W. K. Ng, "Application of stylometry to darkweb forum user identification," in *International Conference on Information and Communications Security*. Springer, 2016, pp. 173–183.
[12] Y. Li, Z. Zhang, Y. Peng, H. Yin, and Q. Xu, "Matching user accounts based on user generated content across social networks," *Future Generation Computer Systems*, vol. 83, pp. 104–115, 2018.
[13] J. Gharibshah, T. C. Li, M. S. Vanrell, A. Castro, K. Pelechrinis, E. E. Papalexakis, and M. Faloutsos, "Inferip: Extracting actionable information from security discussion forums," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 2017, pp. 301–304.
[14] I. Deliu, C. Leichter, and K. Franke, "Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 3648–3656.
[15] B. Biswas, A. Mukhopadhyay, S. Bhattacharjee, A. Kumar, and D. Delen, "A text-mining based cyber-risk assessment and mitigation framework for critical analysis of online hacker forums," *Decision Support Systems*, vol. 152, p. 113651, 2022.
[16] R. Islam, M. O. F. Rokon, E. E. Papalexakis, and M. Faloutsos, "Tenfor: A tensor-based tool to extract interesting events from security forums," in *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2020, pp. 515–522.
[17] A. B. Sayamber and A. M. Dixit, "Malicious url detection and identification," *International Journal of Computer Applications*, vol. 99, no. 17, pp. 17–23, 2014.
[18] H. Tupsamudre, A. K. Singh, and S. Lodha, "Everything is in the name–a url based approach for phishing detection," in *International symposium on cyber security cryptography and machine learning*. Springer, 2019, pp. 231–248.