RURLMAN: Matching Forum Users Across Platforms Using 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*—How can we leverage the URLs posted on online forums to connect forum users with their profiles on other platforms? Most previous studies primarily focus on analyzing textual content and user metadata, paying limited attention to URLs. In this paper, we propose *RURLMAN*, a modular ensemble of methods for leveraging user-posted URLs to connect online forum users with their cross-platform profiles. Our approach has two key features: (a) we focus on user-posted URLs as the key source of information, and (b) we utilize a modular stacked ensemble integrating multiple methods, including string-matching and two ChatGPT capabilities. We show that *RURLMAN* effectively combines the strengths of its component methods, outperforming each individual method with an F1 score of 92.6%. We apply *RURLMAN* in a case study comprising 1.3M URLs posted by 250K forum users across six online security forums and consider URLs to Twitter, Facebook, GitHub, and YouTube. First, we match 30% of the users who shared URLs to these platforms with the corresponding owners of the linked social media profiles. Second, we connect 8% of these users to profiles on multiple platforms. Finally, we identify and analyze "groups" of users based on their posted URLs. To facilitate further research, we will share access to *RURLMAN* and its datasets with the research community.**

I. Introduction

Online forum users often disclose their "identity" on other online platforms through the URLs they share. Currently, there are 100K forums with 500M monthly active users on the internet [1] [2]. Many of these forums harbor malicious hackers responsible for causing major cyber-attacks [3]. We see posted URLs as a missed opportunity to facilitate the disambiguation of users across platforms [4]. At the same time, this could be a "wake-up call" for privacy-conscious users. Note that the URLs that can be used for disambiguation purposes are those that lead to **User-Centric Platforms** where users have personalized accounts and profiles, referred to as **UCPs**. As explained later, we focus on the following UCPs: Twitter, Facebook, GitHub, and YouTube.

ASONAM '23, November 6–9, 2023, Kusadasi, Turkiye

ACM ISBN 979-8-4007-0409-3/23/11.
https://doi.org/10.1145/3625007.3627495

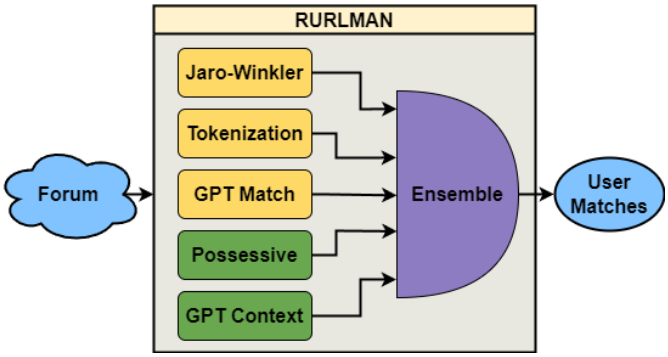

Fig. 1. The *RURLMAN* approach: using posted URLs to connect forum users with their profiles on other platforms. The ensemble combines the output of several string-based methods (yellow) and context-based methods (green) to determine the likelihood of the match.

In this paper, we address the following *user disambiguation* problem: how can we connect forum users with their profiles on other platforms using the URLs they share? The challenge is to develop a method to match a forum user with the person pointed to by the URL, as illustrated in the examples in the following paragraph. The input is one or more online forums, and the output is possibly matching pairs of users across platforms. In expanding the scope of the problem, we want to identify "groups" of forum users across platforms. Here, we define a **group** as a set of users that either: (a) are "members" of an online community, such as a Facebook group, or (b) are aware and interested in a UCP entity, as indicated by them posting a URL to it. In a nutshell, we can see this problem as **establishing a disambiguation seed** of users, which can support subsequent disambiguation steps as we discuss in section VI.

We can consider a practical example to illustrate the challenges of user disambiguation. We have a forum user *subhro* who shares a GitHub repository *subhra74/xdm* which is owned by *subhra74*. Is *subhro* the same person as *subhra74*? While there is some string similarity between the usernames, with a Levenshtein edit distance of 3 and similarity of 62.5%, this alone is insufficient to make a decision. Often, the post that the UCP URL is shared in provides the additional context to help us make an accurate determination. For example, the post could explicitly say "Check out my GitHub repository", which could facilitate the disambiguation. In section VI, we discuss limitations, possible extensions, and potential practical impact

of this research direction.

Most previous studies do not focus on URLs for user disambiguation purposes, instead primarily utilizing textual content analysis and user metadata analysis to reach the same goal. The relevant previous studies can be broadly grouped into three large categories: (a) online user identification methods [5], [6], [7], (b) URL analysis techniques [8], [9], [10], and (c) community discovery studies [11], [12], [13]. We discuss prior work in more detail in section VII.

As our key contribution, we propose *RURLMAN*, a modular ensemble of methods for connecting forum users with their network of off-site profiles. To the best of our knowledge, our work is among the first systematic studies leveraging posted URLs to do user disambiguation across platforms. Our ensemble considers two types of methods: string matching methods and context analysis methods that focus on the text around the posted URL. We also develop a tunable approach to identify groups of users (as defined above) across platforms. We show that *RURLMAN* achieves an F1 score of 92.6% in identifying the social media accounts associated with forum users, outperforming each of its component methods.

As a case study, we apply our methodology on a dataset of six online security forums. The dataset consists of 2.5M posts made by 250K users and span from 2002 to 2022. These forums range from tightly moderated community support forums to disorganized platforms plagued by spam and malicious users [10]. By focusing on security forums, we address the question of whether even tech-savvy users disclose personal information in their URL-sharing activities. We discuss our dataset and its characteristics in detail in section II. We summarize the key results from our case study below.

Observation 1. There is significant URL posting activity in forums. We find that 22% of all forum users post URLs during their participation on the forum, with 4% of them sharing UCP URLs. The majority of users (80%) share only one URL or less, while the most active 1% of users contribute at least 15 URLs throughout their forum engagement. We also find that different forums exhibit distinct preferences for specific UCPs, with some forums heavily favoring a single platform and others having a more balanced distribution.

Observation 2. We can disambiguate the identities of a substantial number of users. We utilize *RURLMAN* to disambiguate the identities of 30% of all users who share UCP URLs, illustrated in Figure 3. Additionally, we connect 8% of these users with their 2 or more external accounts, which provides a more complete picture of a user's digital footprint.

Observation 3. We identify a significant number of cross-platform social groups. We are able to leverage UCP URL sharing activity to identify 556 distinct social groups among forum users. Furthermore, we provide tuning "knobs" that determine the "membership requirement" and can be adjusted to meet the needs of a study, as we describe in section III. These groups often: (a) have large cross-forum membership, and (b) span multiple social media platforms.

Privacy implications. Our work can be seen as a warning for privacy-concerned users: pointing to one's own profiles in other media can reveal one's comprehensive online cross-platform footprint. We discuss this topic further in section VI.

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
MWT	15971	503391	717249	7177	27526
TR	198606	1525107	511321	42327	7853

II. DATASET AND STATISTICS

We provide a description of our dataset, discuss relevant statistics, and describe how we create our ground truth.

A. Dataset. Our work focuses on six online security forums: Offensive Community [14], Ethical Hacker [15], Hack This Site [16], Wilders Security [17], Malware Tips [18], and Tech Republic [19]. For the rest of the paper, we refer to these forums as OC, EH, HTS, WLD, MWT, and TR, respectively. Our data spans a time frame of 21 years, from 2002 to 2022, and contains forums with varying levels of content moderation. Each of these forums contains user-generated content surrounding specific topics. The user-generated content comes in the form of posts that consist of text, images, and links as well as metadata including the user ID, post date, and user title. While the metadata available varies from forum to forum, the structure of posts and threads remains the same throughout. We provide a summary of our dataset in Table I, where *URL Users* refers to users sharing at least one URL.

B. Statistics. In this section we investigate the potential for uncovering user identities on forums by examining the prevalence of user-posted URLs to UCPs that we can use to establish connections with the linked UCP profiles. For the rest of this paper, we refer to a pair of usernames of a forum user and their linked UCP profile as a **user-pair**. Note that user-pairs are not yet confirmed to be a match, but instead represent a potential match of a forum user and a UCP profile. In our investigation, we extract all URLs from forum users' post contents, distinguish the relevant UCP URLs, and model the relationship between the forum and the UCPs. We explain our findings in the following key observations.

Observation 1. There is substantial URL activity on forums. We begin the preliminary investigation with a broad survey of URL activity across the forums in our dataset. Our six forums collectively contain 250K users posting a total of 2.5M URLs. We find 2.6K of these users posting 36.8K URLs to social media platforms. We summarize the key statistics about each forum in our dataset in Table I.

Interestingly, our forums exhibit a wide variety of URL behavioral patterns. Forums TR, MWT, and WLD far exceed the size of OC, HTS, and EH by number of posts and users. In terms of average number of URLs posted per user, OC and MWT far surpass the other forums. We do not find significant correlation between forum size and URL activity per user.

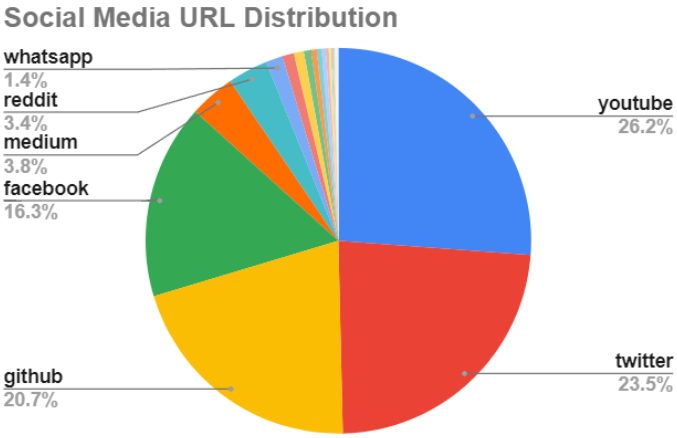

Fig. 2. More than 85% of social media activity can be captured by YouTube, Twitter, GitHub, and Facebook URLs.

Observation 2. Forums exhibit different social media affinities. The intensity and diversity of posted social media URLs varies greatly from forum to forum. In terms of URL diversity, some forums have clear preference for specific UCPs, as with YouTube constituting 96.3% of UCP URLs on TR. On the other hand, other forums are fairly balanced, as with MWT where the most frequently posted UCP is GitHub, constituting 37.1% of UCP URLs on the forum.

In terms of URL intensity, we can see from Table I that MWT has significantly higher social media activity when normalizing for its size compared to all other forums, with an average of 1.7 social media URLs shared per user. WLD, on the other hand, has significantly less relative social media activity, with an average of only 0.02 social media URLs shared per user. This observation serves as an indication that MWT users are more prone to revealing identifiable information from their posted URLs than WLD users.

Observation 3. URL posting activity aligns with the nature of the forum. We investigate URL activity and find that there is good correlation with the overall orientation of the forum. We find that OC, HTS, and EH are spammy and malicious in nature, aligning with their verified gray area orientation [20]. For example, the most common URL category posted in OC is adult entertainment. Upon closer investigation, we discover user "aabee" posting thousands of URLs to adult websites followed by stolen account credentials used to log in to the websites. Additionally, looking at the top URL posting user in OC reveals user "montana" who leaked thousands of proxy servers. We find that the top two users in terms of total posts, "AN0N.PH03N1X" and "tutmoses", advertise their hacking services and provide unsolicited advice on how to perform various cyber attacks, such as stealing Facebook account credentials. This evidence paired with the fact that OC as an online discussion forum is no longer operational leads us to deduce that the platform was sparsely, if at all, moderated and has been taken down by a law enforcement authority.

By contrast, we investigate MWT and find that the top URL posting users include numerous emoticons in each post, which are treated as URLs on the forum and can be considered as benign URL posting activity. Additionally, in TR we find that the top two users in terms of total posts, "oh smeg" and "rob miners", share technical advice and engage in benign conversations. The lack of malicious URL activity found in MWT and TR aligns with their well moderated nature.

C. Creating the ground truth dataset. In the absence of an existing ground truth that labels user-pairs as matches or non-matches, we have to create our own. An effective ground truth dataset needs to be fairly balanced with both matches and non-matches of user-pairs. If we were to select a simple random sample of forum users with posted UCP URLs, there is a low possibility that we would sample a significant amount of matches. We therefore utilize string matching to stratify our sample into high similarity, low similarity, and medium similarity user-pair samples.

Specifically, we construct a ground truth of 150 user-pairs in the following manner. We randomly select three samples of 50 user-pairs with low (0-40%), medium (40-70%), and high (70-100%) string similarity scores between the usernames. We ensured that each selected user appears only once in the ground truth. The similarity score was calculated using Levenshtein edit distance normalized by the longer of the two usernames.

Establishing *D-Manual*. We utilize expert computer scientists as annotators for labeling the ground truth, which we term *D-Manual*. Each annotator is provided with the following information for each of the 150 samples: username of forum user, URL that the user posted, and the post that the URL was shared in. Using this information, each annotator is asked to determine the likelihood that the given URL belongs to the given user, on a scale from 1 to 5. At this point, we check the quality of the annotators using the Cohen's kappa agreement score. We find a kappa score of 67.3% between the annotators, satisfying the requirements as defined by [21]. We then average the annotator scores for each sample, and define the sample as a match if the average score exceeds 3.0.

Establishing *D-Improved*. We created an additional version of the ground truth because in this particular case we were able to find some additional information. In one of our forums, we discovered a thread asking users to post their own social media accounts in order to participate in a giveaway raffle. We reason that the social media accounts posted in this thread belong to the user posting them beyond reasonable doubt. Therefore, any user-pairs from our ground truth originating from this thread may be labeled as a match, even if the post content itself is unintelligible, as in the case of users solely posting their social media URLs. We found 11 such cases in our ground truth where we overruled the decision of the annotators based on this reasoning, and named this new ground truth dataset as *D-Improved*. For full transparency we provide full results from both datasets, although we consider *D-Improved* to be slightly more accurate.

III. Methodology

Our user disambiguation framework consists of two capabilities: (a) we disambiguate user identities, and (b) we

identify cross-platform groups of users, which we describe in the following subsections.

A. Cross-platform user disambiguation. When a user shares a URL pointing to a User-Centric Platform (UCP), we want to determine if the user is the owner of the linked profile. In some cases, the URL points directly to a profile page, and in other cases it leads to a resource (e.g. post, image, repository) owned by a profile. We extract the username of the profile, and create a user-pair from it and the forum username, which we then want to disambiguate as a match or non-match. We use an ensemble approach for this process of matching users across different platforms, which we visually present in Figure 1.

A1. The component methods of our ensemble. Our ensemble consists of two categories of methods: string matching methods and context analysis methods. We provide an explanation of each method below.

String matching methods. The string matching methods take as input a user-pair, consisting of a pair of usernames, and output a classification of either a match or non-match.

(a) Jaro-Winkler matching. We use Jaro-Winkler to determine the similarity of two username strings with emphasis on matching prefixes of the strings. As part of our ensemble, Jaro-Winkler outputs a user-pair as a match if it calculates that the pair has a similarity above a threshold. We provide a default recommended threshold based on our case study in the next section, as well as providing it as an adjustable knob for a practitioner to tune to their specific needs.

(b) Token-based matching. We incorporate a recently developed approach [22] that utilizes token-based matching of usernames. This method emulates human-like interpretation to disambiguate usernames in the following manner. First, it deobfuscates technical and slang usernaming conventions. Then, it splits usernames into meaningful tokens. Finally, all possible token lists generated in the previous step are compared with each other to derive a similarity score. This method is particularly effective for usernames with complex structure and slang conventions. As part of our ensemble, this method classifies a user-pair as a match if it surpasses a similarity score threshold. We use the default threshold here as it was found to work best for online usernames [22], though we also provide it as an adjustable knob for a practitioner to tune to their specific needs.

(c) ChatGPT matching. This method utilizes ChatGPT, found to be effective for classification tasks [23], to attempt to match users with their linked social media accounts. We experimented with several ChatGPT prompts in order to achieve good performance by following best practice guidelines [24] and through manual inspection. In the prompt we provide ChatGPT with the forum username and linked social media username, and we ask it whether the two usernames belong to the same user or not. Here we specifically ask ChatGPT to provide a yes or no answer, and in the future we can experiment with acquiring a similarity score, say from 1 to 5, but this will introduce the additional step of translating the score to an answer. For the rest of this paper, we use the term $\mathbf{GPT_{match}}$ to refer to this method.

Context analysis methods. How can we capture users that use distinctly different usernames across platforms? For example, user "gery79" points to "olger.kapxhiu" on Facebook, but the post says "I did share it on my FB here". In this section, we develop methods that use this contextual information to expand our user matching information domain. The context analysis methods take as input the username of the forum user, linked UCP URL, and the post that the URL was shared in, and output a classification of either a match or non-match.

(a) Possessive word detection. This method inspects the post that a UCP URL was shared in to determine if its user claims ownership of that link. The intuition is that a user may mention owning a linked social media account that has a completely different username than their forum username. This is done by disassembling the post into parts of speech. The method then detects when a possessive pronoun is used within a prespecified threshold of words, or **neighborhood**, before the URL, at which point it asserts that the user claims ownership of the link. We provide the default recommended neighborhood size based on our study in the next section, as well as providing it as an adjustable knob for a practitioner to tune to their specific needs.

(b) ChatGPT context analysis. This method utilizes ChatGPT to investigate the context of a post and determine if a user owns a specified social media account. We provide ChatGPT with a prompt with the following information: username of a forum user, UCP URL that they posted, and the post that the URL was shared in. We conclude the prompt with an instruction to determine whether or not the user owns the social media account linked to by the provided URL. The result is a yes or no answer from ChatGPT, which we designate as a match or non-match output from this component method. For the rest of this paper, we refer to this method as $\mathbf{GPT_{context}}$.

A2. Creating the ensemble model. An ensemble model provides an advantage over its component methods by integrating the strengths of each method while minimizing their weaknesses in performing predictions. We construct *RURLMAN* from methods that offer unique approaches to identify user matches, allowing it to gain increased recall (identifies more user matches) and precision (less false positive matches) over the individual methods. A well-engineered ensemble identifies the strengths and weaknesses of its component methods and combines their individual outputs effectively.

We construct *RURLMAN* as a stacking ensemble model, which uses a meta model to learn the best way to combine the outputs of its base models into a single output. As a whole, our ensemble considers the following information for each possible identification: username of forum user, UCP URL, and the post context that the URL was posted in. As a base layer of the stacking ensemble, each model is given its respective input data from this information, and generates an output. The meta layer of an ensemble model is often simple linear models, as complex models are more prone to overfitting to the current data [25]. We therefore use a logistic regression meta layer, and as we will see in section IV, the model performs sufficiently well.

Note that we evaluated multiple ensembling techniques, which further supported our choice of a logistic regression. First, we considered a simple union of the outputs from each component method. When any method of the ensemble outputs a match for a given user-pair, the entire ensemble outputs a match. This technique yielded nearly perfect recall with low precision. This result shows that a simple union approach leads to low F1 score due to many false positives. Second, we considered a majority voting ensemble, where the ensemble outputs the same as the majority of its component methods. This approach yielded a nearly perfect precision with low recall. This result shows that a majority voting approach leads to low F1 score due to many false negatives that result from some methods' outputs being overruled by the majority. In the future, we can consider additional ensembling models as well.

B. Identifying cross-forum social groups. Another dimension of user disambiguation is identifying the social groups that a user is affiliated with or interested in. Focusing on posted URLs, we want to find users that point to the same entities on UCP platforms. For example, we want to find users that follow the same Twitter user or are members of the same Facebook group. As previously stated, we define the term **group** as a set of users that either: (a) are "members" of an online community, such as a Facebook group, or (b) are aware and interested in a UCP entity, as indicated by them posting a URL to it.

Quantifying the user-group relationship hides some subtleties. This is best explained through an example. In Twitter, there are two granularities of resources: (a) tweets, referred to as **UCP resources**, and (b) Twitter users that post tweets, or **UCP owners**. We can define a group at the granularity of either UCP resources or UCP owners. For example, user A points to a tweet of Twitter user T, while user B points to a different tweet of user T. At the tweet granularity, users A and B are unrelated since they point to different tweets, but at the Twitter users granularity, users A and B are connected as the tweets are from the same user T. All of our UCP platforms have this two-level granularity: YouTube has videos and channels, Facebook has posts and users/pages, and GitHub has repositories and authors. Although we also conducted analysis at the UCP resource level, offering finer level affiliations, we present the results at the UCP owner level due to space limitations.

Next, we want to capture the intensity of a forum user's affinity to a UCP owner. We consider a weighted bipartite graph between forum users and UCP owners. The weights of the graph edges capture the forum user-UCP owner affinity. The weights are composed of two distinct values: (a) w_n, the total number of URLs pointing to the UCP owner, and (b) w_d, the number of distinct URLs pointing to the UCP owner. For example, if user A posts two URLs to YouTube video V_1 from channel C, and another URL to video V_2 from the same channel, the edge connecting user A and channel C is assigned weights of w_n = 3 and w_d = 2. We introduce two thresholds, T_n and T_d, to determine the minimum weight values that will be considered as sufficient interaction. Each group is centered around a unique UCP owner and consists of users with URL activity surpassing these threshold values.

TABLE II
EVALUATION OF EACH INDIVIDUAL ENSEMBLE METHOD FOR *D-Improved*.

Method	Precision	Recall	F1 score
Ensemble	92.4%	**93.0%**	**92.6%**
JaroWinkler	**100.0%**	75.7%	85.6%
Token	98.7%	58.3%	73.2%
GPT_{match}	99.1%	78.3%	87.2%
Possessive	89.3%	13.9%	23.2%
$GPT_{context}$	94.1%	83.5%	88.1%

IV. EVALUATION

We evaluate the effectiveness of *RURLMAN* via three distinct studies: (a) a comparison study against its component methods, (b) an ablation study of each component method, and (c) an agreement study between the component methods. Our evaluation consists of training and evaluating *RURLMAN* against our ground truth datasets using 5-fold cross validation. We report the average F1 scores of the folds for each method which are fairly indicative of the method's performance, as all fold scores fall within 17% of the reported average.

A. Comparison study. In our comparison study, we evaluate the effectiveness of the individual methods within our ensemble in terms of precision, recall, and F1 score. As depicted in Table II for *D-Improved*, *RURLMAN* exhibits a higher F1 score than each of its component methods. We observe the same trend for *D-Manual*. These results show that in situations where both precision and recall are important, *RURLMAN* achieves superior performamnce over each individual method.

Our ensemble achieves significantly higher recall than its component methods. As seen in Table II, $GPT_{context}$ yields the closest recall of any component method, at 83.5%, falling almost 10% below the ensemble recall. This disparity in recall can be attributed to our ensemble successfully utilizing the different information domains of its component methods to recall more results compared to any single component method.

It is worth noting that string similarity-based methods exhibit significantly higher precision than the ensemble but at the cost of lower recall. This behavior can be attributed to the overall context of the situation. The likelihood of a forum user linking to a social media account with a similar username, while not being their own account, is understandably low. Therefore, string similarity methods are generally quite accurate in identifying the match. However, these methods struggle to match accounts with dissimilar usernames, hence the cost in recall. The ensemble achieves the highest recall out of any individual method by leveraging multiple sources of information. Users of *RURLMAN* that wish to prioritize precision can remove lower precision methods from the ensemble.

B. Ablation study. We determine the impact of each component method on the ensemble by ablating them to observe their respective contributions to the overall ensemble performance.

We find that the most significant component methods in *RURLMAN* are $GPT_{context}$ and Jaro-Winkler, in terms of overall F1 score contribution. By ablating $GPT_{context}$ *RURLMAN* loses 2.9% F1 score in the evaluation against *D-Improved*, making

TABLE III
AGREEMENT MATRIX SHOWING THE AGREEMENT OF EACH COMPONENT METHOD FOR *D-Improved*.

Method	JaroWink	Token	GPT_{match}	Possessive	$GPT_{context}$
JaroWink	100%	85%	89%	43%	78%
Token	-	100%	79%	50%	68%
GPT_{match}	-	-	100%	43%	79%
Possessive	-	-	-	100%	39%
$GPT_{context}$	-	-	-	-	100%

it the most significant component method. The second most significant method was Jaro-Winkler, whose ablation resulted in a decrease of 1.0% F1 score. The rest of the methods caused an insignifcant change in F1 score.

In terms of recall, we find that $GPT_{context}$ is again the most significant component method. By ablating $GPT_{context}$ from the ensemble, we observe a decrease of 7.8% in recall. We believe this drop in performance is due to the unique domain that $GPT_{context}$ operates in, where it considers the entire context of the post rather than just the username or possessive pronouns. Other methods with significant recall changes are GPT_{match} and Jaro-Winkler, each causing a decrease of 1.7% recall with their ablation from the ensemble. Possessive matching and token-based matching caused an insignificant change in recall.

Interestingly, ablating $GPT_{context}$ from the ensemble causes a significant increase of 2.6% in precision. This is due to the false positives from $GPT_{context}$, likely caused by the variability in user-generated text making it difficult for ChatGPT to make accurate predictions. The ensemble accepted these false positive results due to their detriment being far outweighed by the benefit in recall that $GPT_{context}$ offers.

C. Agreement study. In this study, we compare the results of *RURLMAN* component methods to assess their level of agreement. In an ensemble, adding models that are very similar to other models offers little to no benefit to the ensemble. An agreement study reveals the uniqueness of each method in the ensemble, and therefore how much it widens the information domain that the ensemble considers. We present our results in an agreement matrix, showing the results in common between each method as a percentage in Table III.

The string matching methods exhibit strong overall agreement of 80-90%. These methods, specifically Jaro-Winkler, Token-based matching, and GPT_{match}, use the same user-pair input to detect matches and might therefore appear repetitive. However, as we saw in the ablation study, removing Jaro-Winkler caused a noticeable detriment to overall performance. We can conclude that the ensemble effectively utilizes the differences between the string matching methods to expand its information domain, even when the methods are quite similar.

The context based methods offer the most unique perspectives to the ensemble, with possessive word detection having the least agreement with other component methods. The specific information domain that possessive detection operates on is best described in an example. We have user "gery79" sharing a UCP URL to profile "olger.kapxhiu" with the post context of "i did share it on my fb" followed by the URL. While every other component method failed to match these users, the possessive word method detected "my" appearing two words before the URL, therefore labeling the user-pair as a match. We suspect that $GPT_{context}$ failed here due to the shorthand, grammatically incorrect language used in the post. The agreement study allows us to identify the niche but strategically important value of this component method.

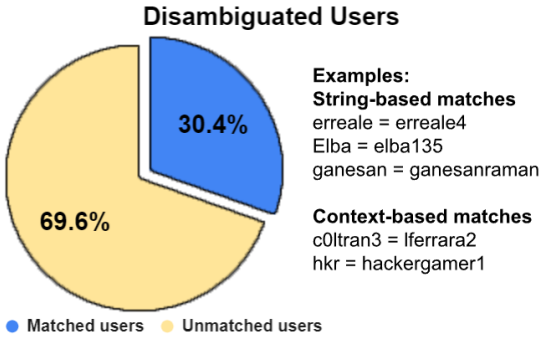

Fig. 3. *RURLMAN* connects 30% of users posting UCP URLs with their cross-platform social media profiles.

V. APPLICATION OF *RURLMAN*

We want to showcase the type of studies and insights that our approach can provide. We apply *RURLMAN* in a case study on our dataset of six online forums containing 2.5M posts made by 250K users, spanning across a 21 year timespan. In this section, we report our results to support the following points: (a) we can disambiguate the identities of select users, and (b) we can identify cross-platform social groups of users.

A. Disambiguating user identities. Using *RURLMAN*, we disambiguate over 30% of UCP-posting users. This percentage consists of 323 distinct users that we matched to a UCP profile out of 1062 UCP-posting users in our dataset. Users in this statistic include any user that was successfully matched to at least one UCP profile. Note that we only focused on four UCPs in this study. In the future, we plan to extend our work to extract user-pairs from more platforms, such as LinkedIn, allowing the disambiguation of more users.

It is particularly interesting that we have matched 26 forum users with more than one UCP. For example, we have MWT user "IObit2013" that we matched with Twitter user "iobitsoft" and Facebook user "iobitsoft", which through manual inspection we found to be a user advertising their own software-selling brand. This multiplatform disambiguation presents a more complete digital footprint for these users, allowing for accurate tracking of their cross-platform activity. We discuss our aim to extend this capability by recursively crawling the matched UCP profiles in section VI.

In our study, we observe varying tendencies among social media platforms in terms of disclosing user identities. Twitter was the most "revealing" platform, with 22% of Twitter URLs pointing to profiles that matched with the forum users that shared them. Facebook follows with 7% of all Facebook URLs revealing the users that shared them. GitHub and YouTube were the least revealing platforms, with less than 2% of these URLs belonging to the users that shared them.

B. Identifying cross-forum user groups. We identify groups of users across the forums in our dataset by analyzing user UCP URL posting activity. We only consider groups that have at least two members from our forums.

We identify a total of 556 groups of users throughout our dataset using the most relaxed thresholds, T_n = 1 and T_d = 1. Each member of these groups shared at least one link to the UCP entity that the group is formed on. We find 271 such groups centered around Twitter, 43 groups centered around Facebook, 23 groups centered around YouTube, and 219 groups centered around GitHub.

Using a stricter threshold, T_n = 2 and T_d = 1, we identify a total of 143 groups of users throughout our dataset. Each member of these groups shared at least two links to the UCP entity that their group is formed on. We find 57 such groups centered around Twitter, 7 groups centered around Facebook, 5 groups centered around YouTube, and 74 groups centered around GitHub.

We investigate groups at very high thresholds, T_n = 10 and T_d = 2, and found a total of six groups, comprising five GitHub groups and one Twitter group. We manually inspected each of these highly dedicated groups to determine their nature. For example, one group centered around Twitter profile "malwaretipscom" and consisted of two MWT staff users. Additionally, the GitHub groups all centered around influential GitHub developers with multiple repositories, including "microsoft" and "adguardteam", and consisted of staff and benign users from multiple forums.

VI. Discussion

We discuss the impact, limitations, and possible extensions of our work.

Practical use: a disambiguation seeding step. *RURLMAN* is best used as a seeding step in the user disambiguation process of a forum. When given a forum to disambiguate its users, we can use *RURLMAN* to create a starting seed of user-pairs and then use additional techniques, such as stylometry comparison, to improve the precision of each match. By using more accurate but computationally expensive techniques, such as [26] [27] [28], on the initial seed from *RURLMAN*, we prevent the combinatorial explosion problem of comparing every possible user-pair across platforms to find matches.

We envision two types of users of *RURLMAN*: (a) security researchers wishing to further study online user behaviors, including privacy concerns, and (b) law enforcement authorities wishing to track malicious users across platforms. Researchers may also use *RURLMAN* to compare and evaluate their tools, as well as to generate a starting seed of potential user matches to use with their own user disambiguation approaches. We will invite and facilitate new methods to be added as components to *RURLMAN* to tailor it to a desired research environment.

Are we sure it is the same user? When *RURLMAN* connects a forum user with an external social media account it does not guarantee that they are the same user. Our approach can only indicate matches that a human would also find. The failure cases for both our approach and a human appear when: (a) a user links to a social media account with a similar username, and (b) a user claims a linked social media account as their own, when it is not. Although these cases are not likely in our experience, they are not impossible. As a result, we see of *RURLMAN* as a starting seed of user matches that can be further filtered out with more computationally expensive methods or even with manual inspection, especially in situations where precision is of paramount importance.

Is our data representative? In any evaluation study, this is a hard and inevitable question. We argue that our data is sufficiently representative. First, we made an effort to consider a diverse set of security forums. Our forums cover a 21 year timespan, with small and massive, closely moderated and unmoderated forums. Second, we consider four social media platforms, comprising both traditional and modern platforms, to ensure that our results are not specific to only one platform.

Ethical considerations. In conducting our work, we followed the best practices according to the ACM-community code of ethical research [29]. We obtained our data from online public forums that did not require login credentials to access. While we mention certain forum users in this work, we will happily obfuscate usernames at the committee's discretion. With a strong commitment to user privacy and responsible data handling, we grant access to *RURLMAN* to researchers and practitioners on a case-by-case basis to ensure adherence to ethical guidelines. Our aim is to mitigate and potentially eliminate any negative impact on online user privacy.

Privacy implications. Our work serves as a warning for privacy-conscious users by shedding light on the risks of pointing to one's own social media accounts. In our case study, we highlight how such behavior may lead to inadvertently disclosing an individual's cross-platform "digital footprint". We strongly urge users to exercise caution and be mindful of the information they disclose by posting URLs online.

Extensions to *RURLMAN*. In the future, we plan to expand the disambiguation capabilities of *RURLMAN* by recursively crawling matched UCP profile pages in order to match additional accounts. For example, a user may post a link to her own GitHub profile, which contains links to her other social media profiles. By crawling the linked profile, we introduce a functionality to *RURLMAN* capable of connecting UCP accounts that are never directly linked to on the forum.

We also plan to improve the precision and increase the confidence of our matches by implementing targeted stylometry analysis of user-pairs. In this extension, we compare the writing style of the forum user with that of posts scraped from the linked UCP account. By using user-pairs as the starting seed for our analysis we avoid the combinatorial explosion problem that typically arises in NLP disambiguation efforts.

VII. Related Work

The specific problem of matching users across platforms using URLs has not been fully focused on yet, however similar problems have been solved tangentially by works that we describe in the following categories.

Alternative user disambiguation studies. An existing study uses stylometry analysis to determine if users on different platforms are the same natural person by comparing their generated content against each other [5]. Other studies aim at linking user generated content to the specific users it originated from [30] [31]. Several efforts utilize user metadata, including user profile description, geolocation, image, and connections to accurately determine users' digital footprint across social platforms [32] [7]. Other efforts combine time stamp metadata and stylometry analysis of user posts to accurately link users across platforms [26] [28] [27].

Identifying malicious URLs. Several recent efforts focus on URLs to extract information from online forums, however their goal seems to be to identify malicious content rather than utilizing URLs for disambiguation purposes, with a specific focus on detecting phishing URLs [33] [8], [10], [9].

Alternative community discovery. There exist several studies that study the dynamics of user interactions and identify implicit communities in online platforms by studying user generated content other than posted URLs [13] [11] [12].

VIII. Conclusion

As our key contribution, we propose *RURLMAN*, a modular ensemble of methods for connecting forum users with their network of off-site profiles, leveraging their posted URLs. Our ensemble considers two types of methods, string matching methods and context analysis methods, stacked under a logistic regression meta layer. We show that *RURLMAN* achieves an F1 score of 92.6% in matching forum users with their social media accounts, outperforming each of its individual component methods. We apply our methodology in a case study on six online security forums comprising 250K users and 2.5M posts spanning 21 years. We report the key findings from our case study: (a) we can disambiguate the identifies of 30% of all users who share URLs to User-Centric Platforms, and (b) we connect 8% of these users to profiles on multiple platforms, and (c) we identify 556 "groups" of users based on their posted URLs. Our work serves as a significant aid to law enforcement agents and researchers in the community that wish to track user activity across platforms.

IX. Acknowledgements

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

References

[1] 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/

[2] Forum software usage distribution on the entire internet. [Online]. Available: https://trends.builtwith.com/cms/forum-software/traffic/Entire-Internet

[3] S. Samtani and H. Chen, "Using social network analysis to identify key hackers for keylogging tools in hacker forums," in *2016 IEEE conference on intelligence and security informatics (ISI)*. IEEE, 2016, pp. 319–321.

[4] B. Treves, M. R. Masud, and M. Faloutsos, "Urlytics: Profiling forum users from their posted urls," in *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2022, pp. 510–513.

[5] S. Vosoughi, H. Zhou, and D. Roy, "Digital stylometry: Linking profiles across social networks," in *Social Informatics: 7th International Conference, SocInfo, Beijing, China*. Springer, 2015.

[6] T. N. Ho and W. K. Ng, "Application of stylometry to darkweb forum user identification," in *2016 ICICS*, K.-Y. Lam, C.-H. Chi, and S. Qing, Eds. Springer International Publishing, 2016.

[7] D. Chatzakou, J. Soler-Company, T. Tsikrika, L. Wanner, S. Vrochidis, and I. Kompatsiaris, "User identity linkage in social media using linguistic and social interaction features," in *Proceedings of the 12th ACM Conference on Web Science*, 2020.

[8] L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, "A novel approach for phishing detection using url-based heuristic," in *2014 ComManTel*, 2014, pp. 298–303.

[9] O. Christou, N. Pitropakis, P. Papadopoulos, S. McKeown, and W. Buchanan, "Phishing url detection through top-level domain analysis: A descriptive approach," *Proceedings of the 6th International Conference on Information Systems Security and Privacy*, 2020.

[10] R. Islam, B. Treves, M. O. Rokon, and M. Faloutsos, "HyperMan: Detecting misbehavior in online forums based on hyperlink posting behavior," in *Social Network Analysis and Mining*, vol. 12, no. 1, 2022.

[11] D. Leprovost, L. Abrouk, and D. Gross-Amblard, "Discovering implicit communities in web forums through ontologies," *Web Intelligence and Agent Systems: An International Journal*, 2012.

[12] T. Chomutare, E. Årsand, L. Fernandez-Luque, J. Lauritzen, and G. Hartvigsen, "Inferring community structure in healthcare forums," *Methods of information in medicine*, 2013.

[13] E. Marin, J. Shakarian, and P. Shakarian, "Mining key-hackers on darkweb forums," *1st ICDIS*, 2018.

[14] Offensive community. [Online]. Available: http://offensivecommunity.net/

[15] Ethical hacker. [Online]. Available: https://www.ethicalhacker.net/

[16] Hack this site. [Online]. Available: https://www.hackthissite.org/

[17] Wilder security. [Online]. Available: http://www.wilderssecurity.com/

[18] Malware tips. [Online]. Available: http://www.malwaretips.com/

[19] Tech republic. [Online]. Available: http://www.techrepublic.com/forums

[20] J. Gharibshah, E. E. Papalexakis, and M. Faloutsos, "REST: A thread embedding approach for identifying and classifying user-specified information in security forums," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2020.

[21] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, p. 276–282, 2012.

[22] M. R. Masud, B. Treves, and M. Faloutsos, "GeekMAN: Geek-oriented username matching across online networks," in *2023 ASONAM*, 2023.

[23] T. Kuzman, N. Ljubešić, and I. Mozetič, "ChatGPT: beginning of an end of manual annotation? use case of automatic genre identification," *arXiv preprint arXiv:2303.03953*, 2023.

[24] Gpt best practices. [Online]. Available: https://platform.openai.com/docs/guides/gpt-best-practices

[25] J. Lever, M. Krzywinski, and N. Altman, "Points of significance: model selection and overfitting," *Nature methods*, vol. 13, 2016.

[26] F. Johansson, L. Kaati, and A. Shrestha, "Timeprints for identifying social media users with multiple aliases," *Security Informatics*, 2015.

[27] 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*, 2018.

[28] F. Johansson, L. Kaati, and A. Shrestha, "Detecting multiple aliases in social media," in *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*, 2013.

[29] *ACM Code of Ethics*. [Online]. Available: https://www.acm.org/code-of-ethics

[30] J. S. Li, L.-C. Chen, J. V. Monaco, P. Singh, and C. C. Tappert, "A comparison of classifiers and features for authorship authentication of social networking messages," *Concurrency and Computation: Practice and Experience*, 2017.

[31] F. Alonso-Fernandez, N. M. S. Belvisi, K. Hernandez-Diaz, N. Muhammad, and J. Bigun, "Writer identification using microblogging texts for social media forensics," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021.

[32] A. Malhotra, L. Totti, W. Meira Jr, P. Kumaraguru, and V. Almeida, "Studying user footprints in different online social networks," in *ASONAM*, 2012.

[33] 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.