Social Network Analysis and Mining (2021) 11:56
https://doi.org/10.1007/s13278-021-00758-8

ORIGINAL ARTICLE

HackerScope: the dynamics of a massive hacker online ecosystem

Risul Islam[1] · **Md Omar Faruk Rokon**[1] · **Ahmad Darki**[1] · **Michalis Faloutsos**[1]

Received: 15 February 2021 / Revised: 10 May 2021 / Accepted: 15 May 2021

Abstract
One would have thought that hackers would be striving to hide from public view, but we find that this is not the case: they have a public online footprint. Apart from online security forums, this footprint appears also in software development platforms, where authors create publicly accessible malware repositories to share and collaborate. With the exception of a few recent efforts, the existence and the dynamics of this community has received surprisingly limited attention. The goal of our work is to analyze this ecosystem of hackers in order to: (a) understand their collaborative patterns and (b) identify and profile its most influential authors. We develop HackerScope, a systematic approach for analyzing the dynamics of this hacker ecosystem. Leveraging our targeted data collection, we conduct an extensive study of 7389 authors of malware repositories on GitHub, which we combine with their activity on four security forums. From a modelling point of view, we study the ecosystem using three network representations: (a) the author-author network, (b) the author-repository network and (c) cross-platform egonets. Our analysis leads to the following key observations: (a) the ecosystem is growing at an accelerating rate as the number of new malware authors per year triples every 2 years, (b) it is highly collaborative, more so than the rest of GitHub authors, and (c) it includes influential and professional hackers. We find 101 authors maintain an online "brand" across GitHub and our online forums. Our study is a significant step towards using public online information for understanding the malicious hacker community.

Keywords GitHub · Hackers · Community · Egonet

1 Introduction

A key thesis of this work is that hacker communities and malware source code are publicly available, which amplifies their abilities and their hacking activities. A strong indication is the emergence of young aspiring hackers, such as the 17-year-old kid from Florida who reportedly was the mastermind behind the recent hacking of Twitter (Aaron Holmes 2020). We argue that the security community is paying relatively little attention to these online malicious communities. We see this as a missed opportunity. On the one hand, the hacker community is fairly wide encompassing curious teenagers, aspiring hackers and professional criminals. On the other hand, the hackers are surprisingly bold in leaving a digital footprint, if one looks at the right places in the Internet. For example, there are various online forums, where hackers not only share information, but they also boast of their successes.

The problem we tackle here is the need to analyze and model the ecosystem of malicious hackers based on their online footprint. The input is the online activities of these hackers, and the goal is to answer the following questions: (a) do these hackers work in groups or alone, and (b) who are the most influential hackers? Here, we consider two types of platforms that hackers frequent: (a) software archives and (b) online security forums. It turns out that popular and public software archives, such as GitHub harbour malware authors, who create publicly accessible malware repositories (Rokon et al. 2020). Furthermore, online forums have recently emerged as marketplaces and information hubs of malicious activities (Gharibshah et al. 2020; Portnoff et al. 2017). In the rest of this paper, we will use the term hacker

✉ Risul Islam
risla002@ucr.edu

Md Omar Faruk Rokon
mroko001@ucr.edu

Ahmad Darki
adark001@ucr.edu

Michalis Faloutsos
michalis@cs.ucr.edu

[1] University of California Riverside, Riverside, USA

Published online: 14 June 2021

to refer to actors who develop and use software of malicious intent. We will also use the term *hackers* and *malware authors* interchangeably, although some malware authors may not have malicious intent.

There is limited work for the problem as defined above. First, we are not aware of a study that systematically profiles the dynamics of the online hacker ecosystem, and especially one considering software archives. Most of the previous efforts on GitHub follow a software-centric view or study GitHub at large without focusing on malware (Calleja et al. 2016, 2018; Blincoe et al. 2016). Most of the previous works on online forums focus on identifying emerging topics and threats (Gharibshah et al. 2020; Portnoff et al. 2017). Other efforts report malware activity, focusing on hacking events, and much less, if at all, on the ecosystem of hackers (Sapienza et al. 2017, 2018). We elaborate on previous works in Sect. 8.

We propose HackerScope, a systematic approach for modelling the ecosystem of malware authors by analyzing their online footprint. We start with an extensive analysis of malware authors on GitHub, as this is a significantly less-studied space. We then use security forums to find more information about these authors. From an algorithmic point of view, we use three network representations: (a) the author-author network, (b) the author-repository network and (c) cross-platform egonets, which we explain later. In addition, we use some basic Natural Language Processing techniques, which we intend to develop further in the future.

We apply and evaluate our approach using 7389 malware authors on GitHub over the span of 11 years and leverage the activity on four security forums in the grey area between white-hat and black-hat security. GitHub is arguably the largest repository with roughly 30 million public repositories, while, appropriately fine-tuned, our approach can be used on other software archives. Our approach encompasses four research thrusts, which identify and model: (a) statistics and trends, (b) communities of hackers and their dynamics, (c) influential hackers and (d) hacker profiles across different online platforms. For the latter type, we show the collaborators of hackers as captured by the cross-platform egonets spanning GitHub and security forums in Fig. 1. Our key results are summarized in the following points.

a. The ecosystem is growing at an accelerating rate The number of new malware authors on GitHub is roughly tripling every two years. This alarming trend points to the importance of monitoring this ecosystem.

b. The ecosystem is highly collaborative We find 513 collaboration communities on GitHub with high cohesiveness (*Modularity Score* within [0.65-0.78]), including many large communities with hundreds of users. The malware community is very collaborative: a malware repository is forked *four times* more compared to a regular GitHub repository.

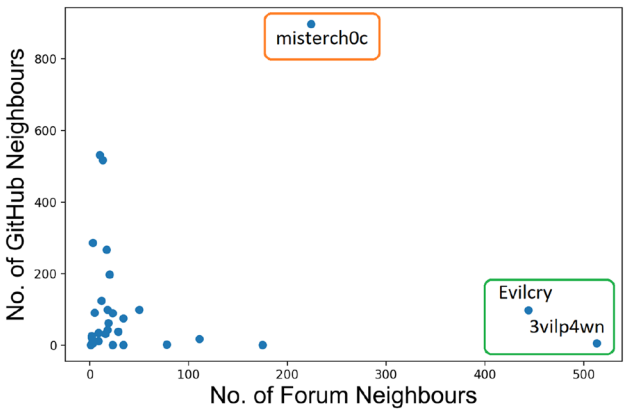

Fig. 1 Profiling hackers across platforms using our cross-platform egonet: the scatterplot of the number of neighbours on GitHub versus those on security forums for 30 malware authors as captured in our cross-platform egonet

c. We identify a group of 1.7% of influential authors We develop a systematic approach to determine the influence among malware authors. Our novelty lies in: (a) considering many types of interactions and (b) capturing the network-wide influence of an author. We find a core group of 1.7% of the malware authors, who are responsible for: (a) generating influential repositories and (b) providing the social backbone of the malware community.

d. We identify professional hackers in the ecosystem We find that 101 authors are professional *malicious* hackers. Going across platforms, we find GitHub authors who are quite active on our security forums. We show the evidence that these are professional hackers, who are building an online "brand". For example, user *3vilp4wn* is the author of a keylogger repository on GitHub, which he promotes in the *HackThisSite* forum using the same username (shown at bottom right in Fig. 1).

Our work in perspective The proposed work is part of an ambitious goal: we want to model the Internet hacker ecosystem at large as it manifests itself across platforms. Our initial results are promising: a) the hackers seem to want to establish a brand; hence, they want to be visible, and b) a cross-platform study is possible, as some authors maintain the same login name. Our systematic approach here constitutes a building block towards the ultimate goal. With appropriate follow-up work, achieving this goal can have a huge practical impact: security analysts could prepare for emerging threats, anticipate malicious activity and identify their perpetrators.

Open sourcing for maximal impact. We use Python v3.6.2 packages to implement all the modules of HackerScope. We

intend to make our datasets and tools public for research purposes.

2 Background and data

Our work focuses on GitHub, the largest software archives with roughly 30 million public repositories and uses data from online forums (security and gaming forums). Although GitHub policies do not allow malware, authors do not seem to abide by them.

A. GitHub data GitHub platform enables software developers to create software repositories in order to store, share and collaborate on projects and provides many social-network-type functions.

We define some basic terminology here. We use the term *author* to describe a GitHub user who has created at least one repository. A *malware repository* contains malicious software, and a *malware author* owns at least one such repository. Users can *star, watch* and *fork* other *malware repositories*. *Forking* means creating a clone of another repository. A forked repository is sometimes merged back with the original parent repository, and we call this a *contribution*. Users can also *comment* by providing suggestions and feedback to other authors' repositories.

We use a dataset of 7389 malware authors and their related 8644 malware repositories, which were identified among 97K repositories in our prior work (Rokon et al. 2020). This is arguably the largest malware archive of its kind with repositories spanning roughly 11 years. These repositories have been identified as malicious with a very high precision (89%). Note that the queries with the GitHub API, which were used in the data collection, return primary or non-forked repositories. A discussion on the process, accuracy and validity of the dataset can be found in the original study (Rokon et al. 2020).

For each malware author in our dataset, we have the following information: (a) the list of the malware repositories created by her, and (b) the list of followers. For each malware repository, we have the lists of users, who: (a) star, (b) watch, (c) fork, (d) comment or (e) contribute to the repository.

Repository metadata Each repository is also associated with a set of user-generated fields, such as title, readme file, description. We can use this *metadata* to extract information about the repository. We leverage our earlier work where we discuss the processing of this metadata in more detail (Rokon et al. 2020).

For a given repository, a security expert would want to know: (a) the type of malware (e.g. ransomware and keylogger), and (b) the target platform (e.g. Linux and Windows). For this, we define two sets of keywords: (a) 13 types of malware, S_1 and (b) 6 types of target platforms, and S_2. Figure 7 provides a visual list of these two sets of keywords.

We define the Repository Keyword Set, W_r, for repository r, as a set consisting of the keyword sets S_1 and S_2 that are present in its metadata. Clearly, one can extend and refine these keyword sets, to provide additional information, such as the programming language in use, which we will consider in the future. Note that our earlier work provides evidence that using this metadata as we do here can provide fairly accurate and useful information (Rokon et al. 2020).

B. Security forum data We also utilize data that we collect from four security forums: Wilders Security, Offensive Community, Hack This Site and Ethical Hackers (Online Forums 2021). In these forums, users initiate discussion threads in which other interested users can post to share their opinion. Each tuple in our dataset contains the following information: forum ID, thread ID, post ID, username, and post content. We provide a brief description of our forums below, and an overview of key numbers in Table 1.

a. OffensiveCommunity (OC) As the name suggests, this forum contains "offensive security"-related threads, namely, breaking into systems. Many posts consist of step-by-step instructions on how to compromise systems, and advertise hacking tools and services.

b. HackThisSite (HTS) As the name suggests, this forum has also an attacking orientation. There are threads that explain how to break into websites and systems, but there are also more general discussions on cyber-security.

c. EthicalHackers (EH) This forum seems to consist mostly of "white-hat" hackers, as its name suggests. However, there are many threads with malicious intentions in this forum.

d. WildersSecurity (WS) The threads in this forum fall in the grey area, discussing both "black-hat" and "white-hat" skills.

Table 1 Basic statistics of our datasets

Dataset	User	Thread/repository	Post	Active days
Offensive community	5412	3214	23918	1239
Ethical hacker	5482	3290	22434	1175
Hack this site	2970	2740	20116	982
Wilder security	3343	3741	15121	777
MPGH	37001	49343	100001	289
GitHub	7389	8644	–	2225

C. Gaming forum dataset We consider an online gaming forum, Multi-Player Gaming and Hacking Cheats (MPGH) (Online Forums 2021). MPGH is one of the largest online gaming communities with millions of discussions regarding different insider tricks, cheats, strategy and group formation for different online games. The dataset was collected for 2018 and contains 100K comments of 37K users (Pastrana et al. 2018).

3 Our approach

We have an ambitious vision for our approach, which we plan to release as a software platform. We provide a brief overview in Fig. 2. In this paper, we will elaborate on the four analysis modules: (a) a statistics and trends module, which provides the landscape of primary behaviours of the ecosystem (Sect. 4), (b) a community analysis module, which identifies and profiles communities of collaboration (Sect. 6), (c) an influence analysis module, which defines and calculates the significance of authors (Sect. 5) and (d) cross-platform analysis module (Sect. 7).

In addition, our approach also includes: a data collection module, which aggregates, cleans and preprocesses the raw information; a control centre module; and a reporting module. These modules are not equally developed, while at the same time, we could not provide all the types of results that we have available due to space limitations.

Below, we highlight some interesting or novel aspects of our approach, which are often cutting across several modules.

a. Synthesizing multi-source data Our approach focuses on data for authors from GitHub and combines it with additional data from security forums, and Internet searches.

b. Defining appropriate features As we already saw, the authors and the repositories have a very rich set of interactions. We have primary (measured directly) and secondary (derived from the primary) features, which need to be determined carefully to capture effectively the dynamics of the ecosystem. These interactions go beyond a simple "friend" relationship of other social media.

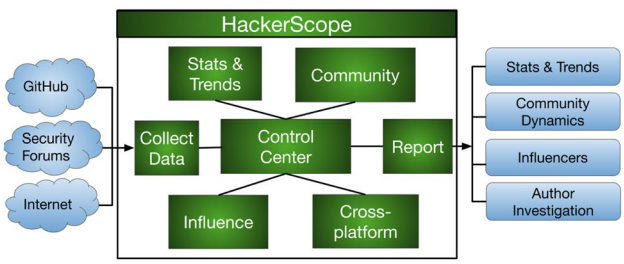

Fig. 2 The overview of our approach highlighting the key functions

c. Modelling the dynamics We use three network representations to capture the rich interactions and relationships among authors and repositories. The network representations include: (a) the author-author network, (b) the author-repository network and (c) cross-platform egonets.

d. Reporting behaviours The goal is to provide intuitive and actionable information in an appealing and ideally interactive fashion. The results in this paper provide an indication of some initial plots and tables that our approach will provide to the end user, who could be a researcher or a security analyst.

4 Statistics and trends

This section describes the functionality of the *statistics and trends* module of our approach, whose intention is to provide a basic understanding of author behaviours.

A. Basic distributions of malware authors. We study the complementary cumulative distribution function (CCDF) of three metrics: (a) the number of repositories created, (b) the number of followers and (c) sum of the number of forks across all the malware repositories of the author. As expected all distributions are skewed, but the plots are omitted due to space constraints. First, we find that 15 authors are contributing roughly 5% of all malware repositories, while 99% of all authors have created less than 5 repositories each. Second, we find that 3% (221) of the authors have more than 300 followers each, while 70% of the authors have less than 16 followers. Finally, examining the total number of forks per author, we find that 3% (221) of the authors have their repositories forked more than 150, while 43% of authors encounter at least one fork.

B. Forking behaviour: Malware repositories are forked four times more than the average repository. Malware repositories are more aggressively forked, which is an indication of the higher collaboration in the ecosystem. First, we find that a malware repository is forked 4.01 times on average, while a regular GitHub repository is forked 0.9 times, as reported in previous studies (Jiang et al. 2017). Second, we want to see whether this is due to a few popular repositories, but this is not the case. We find that 39% of the malware repositories are forked at least once, while this is true for only 14% for general repositories (Jiang et al. 2017).

C. Trends "How fast is this ecosystem growing?" To answer the question, we plot the number of new malware authors per

year in Fig. 3. We consider that an author joins the ecosystem at the time that they create their first malware repository in our database.

a. The number of new malware authors almost triples every two years We plot the new malware authors per year in Fig. 3. We observe an increase from 238 malware authors in 2012 to 596 authors in 2014 and to 1448 authors in 2016. We also observe a steep 62% increase from 2015 to 2016. This trend is interesting and alarming at the same time.

b. The number of new malware repositories more than triples every four years Echoing the growth of the authors, the number of repositories is also increasing superlinearly. In the future, we plan to study the trends of malware in terms of both types of malware and its target platform.

5 Identifying influential authors

To understand the dynamics of the ecosystem, we want to answer the following question: *"Who are the most influential authors?"* The functionality in this section is part of the *influence analysis* module of Fig. 2.

A. HackerScore: Identifying influential authors We argue that finding influential authors presents several challenges. First, there are many different activities and interactions, such as creating repositories, commenting, following other authors and being followed by other authors. Second, we can consider two types of actions: (a) creating influential artefacts, (b) observing and engaging with other people and artefacts. Furthermore, the distinction is not always clear. For example, forking a repository creates a new, but derivative, repository.

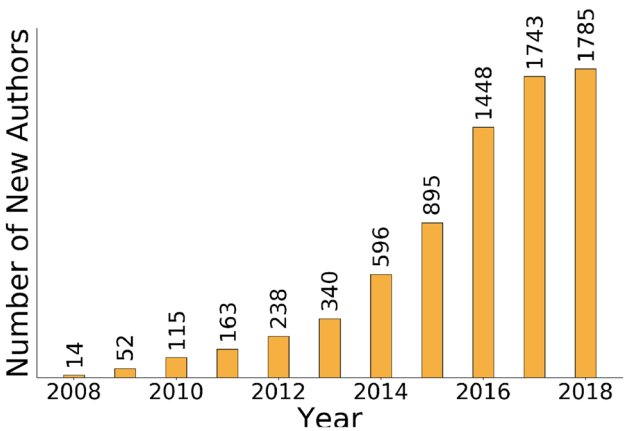

Fig. 3 New malware authors in the ecosystem per year

To address the above challenges, we take socially aware approach to influence: creating a few influential repositories is more important than creating many non-influential repositories. We discuss how we model and calculate this influence below.

The author-author graph (AA) We create the author-author network to capture the network-wide interaction among authors. We define a weighted labelled multi-digraph: $G(V, E, W, L_e)$ where V is the malware author set, E is the set of edges, W is the weight set, and L_e is the set of labels that an edge e can be associated with. These labels correspond to different types of relationships between authors. Here we opted to consider only malware authors in the graph to raise the bar for being part of the hacker community.

The types of interactions We consider four types of relationships between authors here. A directed edge (u, v) from author u to v can be (i) a follower edge: when u follows v, (ii) a fork edge: when u forks a repository of v, (iii) a contribution edge: u contributes code in a repository of v and (iv) a comment edge: u comments in a repository of v. These relationships capture the most substantial author-level interactions.

The multi-graph challenge and weight calibration Our graph consists of different types of edges, which represent different relationships that we want to consider in tandem. The challenge is that the relationships have significantly different distributions, which can give an unfair advantage or eliminate the importance of a relationship. For example, contribution activities are rarer compared to following, but one can argue that a contribution to a repository is a more meaningful relationship and it should be given appropriate weight.

For fairness, we make the weight of a type of edge inversely proportional to a measure of its relative frequency. In detail, we calculate the average degree d_{type} for each *type* of edge: follower, fork, contribution and comment from the subgraph containing only that type of edges from the AA graph. We find the following average degrees: $d_{follower} = 12.21$, $d_{fork} = 4.67$, $d_{contribution} = 0.53$ and $d_{comment} = 0.49$. We normalize these average degrees using the minimum average degree ($d_{min} = 0.49$), and we get the *inverse* of this value as the weight for that edge, namely, d_{min}/d_{type}. This way, we set the following weights: $w = 0.04$ for a following edge, $w = 0.1$ for a forking edge and $w = 1$ for a commenting or a contribution edge. This enables us to consider each relationship type more fairly and meaningfully.

We propose a socially aware and integrated approach to combine all the author activities in a single framework. First, we identify and define two roles in the ecosystem: (a) producers, who create influential malware repositories, and (b) connectors, who enhance the community by engaging with

influential malware authors and repositories. To calculate the roles of the malware authors, we first model the interaction among authors in the *AA* graph described above. Next, we apply our algorithm, a customized version of a weighted hyperlink-induced topic search (WHITS) algorithm modified to handle the multiple types of relationships between authors. We discuss the related algorithms in Sect. 8.

Calculating the HackerScore We associate each node u with two values: (a) a producer HackerScore value, PHS_u, and (b) connector HackerScore value, CHS_u. Let $w(u, v)$ be the weight of edge (u, v) based on its label, as discussed above.

The algorithm iterative refines the producer and connector values until it converges. We, now, elaborate on the steps. First, PHS_u and CHS_u are initialized to 1. During the iterative step, the algorithm updates the values as follows: (i) for all v pointing to u: $PHS_u = \sum_v w(v, u) * CHS_v$, or zero in the absence of such edges, (ii) for all z pointed by u: $CHS_u = \sum_z w(u, z) * PHS_z$, or zero in the absence of such edges, and (iii) we normalize PHS_u and CHS_u, so that $\sum_u PHS_u = \sum_u CHS_u = 1$. For the convergence, we set a tolerance threshold of 10^{-9}. We provide the evidence of convergence in Fig. 4 where we plot the number of nodes changed their values per iteration. We find that after 450 iterations, only 9 out of 7389 nodes changed their values and these values are within the tolerance ($\approx 10^{-11}$).

After the convergence, we obtain reliable two HackerScore values for each author.

Identifying influential malware authors In Fig. 5, we plot the Connector HackerScore versus Producer HackerScore for our malware authors. Separately, we identify "knees" in the individual distributions of each score at PHS = 0.00215 and CHS = 0.0029 indicated by the red dotted lines. This way, we observe four regions defined by the combination of low

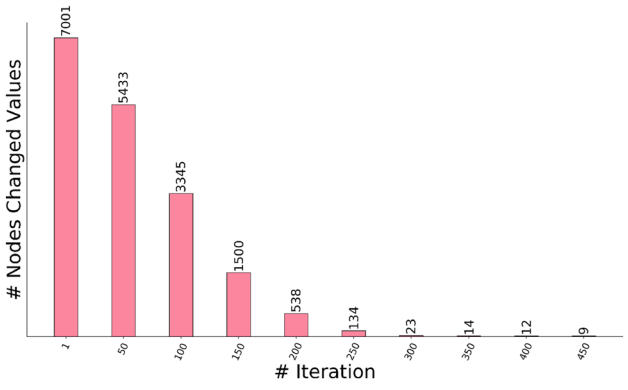

Fig. 4 Customized WHITS algorithm converges when iteration increases. The number of nodes that change their HackerScore values decrease and finally stabilize when number of iteration increases

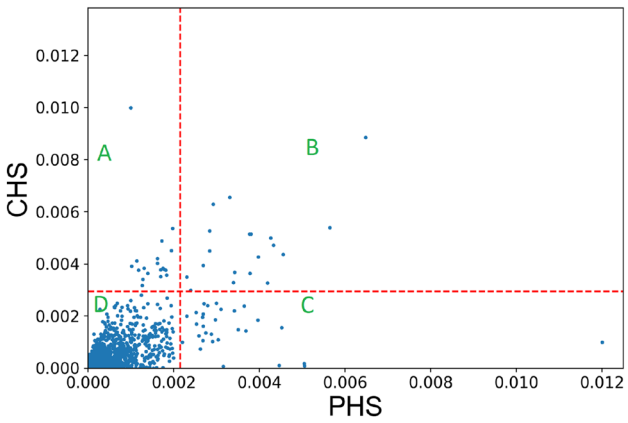

Fig. 5 The scatterplot of the Connector HackerScore vs. Producer HackerScore for the malware authors in our GitHub dataset

and high values for PHS and CHS values which shows if an author is influential as producer or connector.

A few authors (1.7%) drive the community The three regions of influence together consist of 128 malware authors (1.7%). The breakdown of the region size is fairly even: Region A of mostly connector authors devoted to connect the malware community is 0.6%, Region C of the influential producers who are the originator of the malware resources is 0.7%, and Region B of dual influence is 0.4%. We use the term highly influential group (HIG) to refer to this group of authors.

We provide a profile overview of the two most influential authors per region in Table 2. The most influential author of Region C is *cyberthreats*, with the highest PHS (0.012) and 336 malware repositories. She gained a huge following by creating all her repositories of assembly code malware on 16 Feb 2016. The top connector author from Region A is *critics* with a CHS score of 0.01, which stems from her 446 comments across 301 repositories. The top malware author from Region B is *D4Vince* for his dual role in producing credential reuse tools with 7 repositories and 165 comments and 187 contributions.

Validate the influential authors using TenFor We verify the identified influential authors using TenFor tool (Islam et al. 2020b). TenFor takes a tensor as input and reports a bunch of important clusters of authors and dominant authors from each cluster leveraging tensor decomposition. As input, we construct a 3D tensor for GitHub dataset where each element, $T(i, j, k)$, of the input tensor captures the interaction (in terms of the total number of create, fork, comment and contribution performed) between: (a) author i, (b) repository j, (c) per week k. Applying HackerScope on this tensor, we extract a total of 22 clusters. Setting $k = 5$ in TenFor, we get a total of 22*5 = 110 dominant authors. Interestingly, out of this 110 authors, our customized WHITS algorithm is

Table 2 The profiles of the two most influential malware authors from each region A, B and C

Name	PHS	CHS	Repositories	Followers	Forks	Comments	Contributors
Cyberthreats	**0.012**	0.001	336	1013	778	13	2
ytisf	**0.005**	10^{-6}	12	606	1412	4	1
Critics	0.001	**0.01**	6	396	83	446	301
Samyk	0.0018	**0.0058**	2	554	125	176	209
D4Vince	**0.0066**	**0.0082**	7	608	499	165	187
n1nj4sec	**0.0058**	**0.0052**	8	876	1391	64	79

The bold score reflects the author's primary role (producer or consumer) in GitHub

also able to extract 97 authors. That means 98 out of driver 128 (1.7%) authors are caught by TenFor as well. This finding validates the fact that our method does indeed capture influential authors.

The importance of socially aware significance We argue that our socially aware definition of significance provides more meaningful results than simply taking the top-ranked users in any primary metric in isolation. First, the two scores capture different aspects of influence: they can differ by orders of magnitude as is the case with *cyberthreats* and *ytisf*. Second, our scores capture a combined network-wide influence that each primary metric could miss. For example, our most influential producers do not always own many malware repositories. Malware author *D4vince* and *n1nj4sec*, mentioned in Table 2, have single-digit repositories (7 and 8, respectively) and yet are two of the top producers. On the other hand, author *kaist-is521* is ranked way below than *n1nj4sec* in terms of HackerScore (PHS =0.0001 and CHS =0.00013), although she has 18 malware repositories.

B. Reciprocity of interactions We want to understand better the nature of the author interactions here.

"Is the influence among malware authors reciprocal?" The answer is negative: the relationships are not reciprocal, which is in stark contrast to the reciprocal relationships in other social media like Twitter and Facebook (Weng et al. 2010). We consider a total of six relationships: following, forking, commenting, contributing, watching and starring relationships. We define the Reciprocity Index for relationship x, RI_x, to be the ratio of reciprocal relationships over the pairs of authors with that type of relationship (unilateral or mutual) in the author-author network.

We find that the reciprocity is low and less than 7.3% for all the relationships in question.

By contrast, reciprocity is often above 70% in social media, like Facebook or Twitter (Weng et al. 2010). These social media mirror personal relationships and have an etiquette of conduct. We conjecture that the lower reciprocity on GitHub is due to its utilitarian orientation: following an author stems from a professional interest.

6 Community analysis

This section describes the functionality of the *community analysis* module, whose goal is to identify the communities of collaboration among the malware authors on GitHub.

A. Identifying collaboration communities We quantify the collaborative nature of the malware authors as follows.

The author-repository graph (AR) We define the author-repository graph to be an undirected bipartite graph, $G = (A, R, E)$, where A is the set of malware authors and R is the set of malware repositories. An edge $(u, r) \in E$ exists, if author u: (a) creates, (b) stars, (c) forks, (d) watches, (e) comments or (f) contributes to repository r.

Identifying bipartite communities To identify communities, we employ a greedy modularity maximization algorithm modified for bipartite graphs as we discuss in our related work.

We find a total of 513 communities spanning a wide range of sizes as shown in Fig. 6. The size of the communities

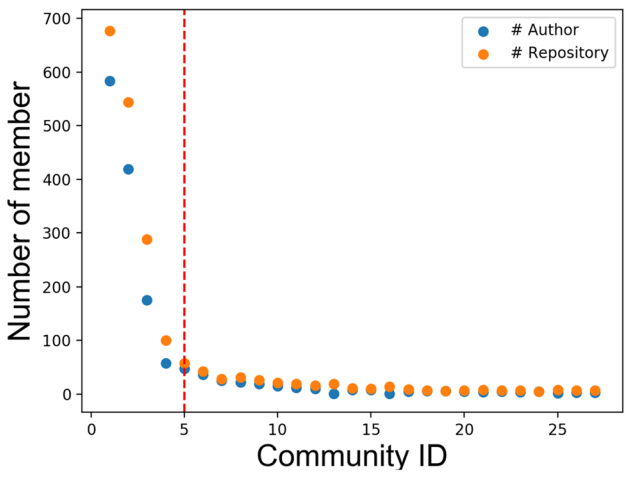

Fig. 6 The distribution of the number of authors and repositories for the 27 largest communities in the order of community size

follows skewed distribution. In Fig. 6, we plot the number of malware authors and repositories per community in order of decreasing community size (defined as the sum of authors and repositories). We find that 90% of communities have less than 14 authors and repositories. We also see a fairly sharp knee in the plot at the fifth community, as shown by the vertical line.

B. Profiling the communities A full investigation of the purpose, evolution and internal structure of each community could be a research topic in its own right. Here, we only provide an initial investigation around the following three questions.

a. How cohesive are our communities? We report the Modularity Score (MS_C), which quantifies the cohesiveness of a community C.

The MS_C is defined as follows: $MS_C = \frac{n_C(E)}{N_C(E)}$, where $n_C(E)$ is the total number of edges and $N_C(E)$ is the number of all possible edges in community C (if the community was a bipartite clique).

Overall, our communities are highly cohesive: 82.8% (425) of the communities have a *Modularity Score* $MS_C \geq 50\%$, which means that more than half of all possible edges within the community exist. Interestingly, the largest communities exhibit strong cohesiveness. In Table 3, we present a high-level profile of the five largest communities which have a Modularity Score of 0.65-0.78, which is indicative of tightly connected communities.

b. Who are the community leaders? We want to identify the influential authors as part of profiling a community. We identify the top two most influential producers and connectors per community using the HackerScore from Sect. 5. This leads us to a group of MFok:144 leaders of the communities of size of at least 20 authors. We find MFok:81% of these community leaders are part of the Highly Influential Group (HIG) of authors. This suggests that the HIG authors are indeed driving forces for the ecosystem. In the future, we intend to investigate in more depth the influence and dynamics of each community.

c. What is the focus of each community in terms of platform and malware type? A security expert would want to know the main type of malware (e.g. ransomware) and the target platform (e.g. Linux) of a community. We use the Repository Keyword Set, W_r, information of a repository r, as we defined in Sect. 2, and we use it to characterize the community.

One way to quantify the importance of a keyword for a community is to measure the number of repositories, for which that keyword appears at least once. In detail, we use the Strength Of Presence *(SOP)* metric, which we define as follows. For a community C with a set of R repositories, we define k_i to be the number of repositories, in which keyword i appears in the metadata W_r for repository r at least once for all repositories $r \in R$. We define the SOP_i of keyword i from keyword set S as follows: $SOP_i = \frac{k_i}{\sum_{j \in S} k_j}$. In Table 3, we show the most dominant keywords from malware types and platforms sets for each community and the related SOP scores.

We can also use the *SOP* to visualize the keywords as a word-cloud. A word-cloud is a more immediate, appealing and visceral way to display the information. In Fig. 7, we show the word-cloud for the third largest community, which is dominated by *ransomware* malware and targets *Windows* platforms. Not only we see the main words stand out, but their relative size conveys their dominance over the other words more viscerally than a lengthy table of numbers.

Fig. 7 The word-cloud for the malware types and platforms keywords for the third largest community: Ransomware and Windows dominate

Table 3 High-level profile of the five largest communities of malware authors and malware repositories

ID	Authors	Repositories	Modularity score	Dominant platforms	SOP	Dominant types	SOP
1	584	677	0.65	Linux	0.32	Keylogger	0.29
2	419	544	0.67	Windows	0.26	Virus	0.31
3	175	288	0.73	Windows	0.65	Ransomware	0.44
4	57	100	0.78	Linux	0.43	Spyware	0.43
5	47	57	0.71	Mac	0.33	Trojan	0.22

We present the results of this type of profiling for the largest communities in Table 3, which we also discuss below.

We find the largest community of 584 malware authors and 677 malware repositories having Linux ($SOP = 0.32$) and keylogger ($SOP = 0.29$) as the dominant platform and malware type. Interestingly, we find that 49 of the top 100 most prolific (in terms of the number of repositories created) authors are in this community. Upon closer investigation, we find that 11 out of the 15 authors with the highest degree in the subgraph of this community are keylogger developers.

The third-largest community consists of 175 malware authors and 288 malware repositories and revolves around Ransomware ($SOP = 0.65$) and Windows platform ($SOP = 0.44$). For reference, we present the word-cloud of the malware types and platforms based on the *SOP* score in Fig. 7 for this community which exhibits that Ransomware and Windows possess the highest *SOP* scores.

Finally, the fourth largest community (57 authors, 100 repositories) is the most tightly connected ($MS = 0.78$) and it revolves around the development of attack tools for Kali Linux. Upon closer inspection, we find that 15 of the top 25 authors (based on node degrees) form an approximate bipartite clique with 5 repositories. This group developed *WiFiPhisher* in 2016, a Linux-based python phishing tool (Sophron 2014), which has been used for both good and evil (Cybersec 2018).

The above are indicative of the potential information that we could extract from these malware repositories. In the future, we intend to: (a) extract more detailed textual information from each community and (b) study the evolution and dynamics of these communities over time.

7 Author investigation

"Who are these malware authors?" To answer this question, we go across platforms to online forums and leverage our datasets from several security and gaming forums. The functions described here are part of the *author investigation* module of Fig. 2.

a. Malware authors strive for an online "brand" and usernames seem persistent across online platforms We find that many malware authors use the same username consistently across many online platforms, such as security and gaming forums, possibly in pursuit of a reputation.

We identify 101 malware authors who are active in one of our five security and gaming forums: 71 in MPGH, 12 in Wilders Security, 6 in Ethical Hacker, 4 in Offensive Community and 8 in Hack This Site (Online Forums 2021). We argue that some of these usernames correspond to the same users based on the following two observations.

First, we find significant overlap in the interests of the cross-platform usernames. For example, usernames *int3grate* and *jedisct1* show interest in ransomware in both platforms, while *3vilp4wn* advertises her keylogger malware (github.com/ 3vilp4wn/CryptLog) in the forum. Second, these usernames are fairly uncommon, which increases the likelihood of belonging to the same person. For example, the top ten results from Internet searching for the username of author *Misterch0c* return nine hacker-related sites and a twitter account with a different handle but claimed by Misterch0c. Note that not all the malware authors or repositories have a malicious purpose. For instance, the project "Empire" (EmpireProject 2018) by *xorrior* was created as an offensive tool to stress-test the security of systems. However, it has recently been used by the state-sponsored hacking group *Deep Panda* (Mitre 2019). In general, offensive security tools contribute to the power of the malware ecosystem irrespective of the intention of its creator.

b. Modelling the cross-platform interactions We propose to study the cross-platform interactions between GitHub and online forums (security and gaming) as a promising research direction that can bridge two domains: software repositories and online forums.

We define the cross-platform egonet of a user as one that consists of her egonets from the two platforms as shown in Fig. 8. The forum egonet captures the interaction of the users that post on the same threads, while we leverage the author-author network to define the GitHub egonet.

The value of cross-platform analysis Using the cross-platform egonet as a basis, we can model the cross-platforms user dynamics, and more specifically, we can: (a) identify common "friends" between the ego-nets, (b) find the topics of interest and activities in each egonet and (c) model information flow and influences across platforms. In Fig. 1, we visualize the activity of a cross-platform (GitHub-Security forum) user by comparing the number of users on each side

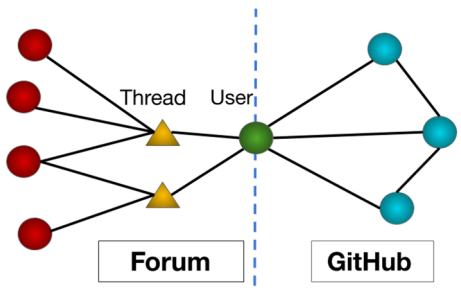

Fig. 8 A cross-platform egonet: capturing the neighbours of both the online forum and GitHub

Table 4 Profiles of eight cross-platform users

Name	Forum	Posts in forum	Collaborators in forum	Malware repo	Follower	Fork	Collaborators in GitHub	Repository content	Internet-wide reputation
misterc0c	WS	7	224	7	749	81	898	Cracked malware code	Self-declared hacker
3vilp4wn	HTS	103	513	1	0	1	6	Python keylogger	Keylogger developer
fahimagsi	OC	73	175	1	1	0	1	Backdoor	Famous hacker
Evilcry	EH	18	444	2	89	15	98	Botnet and ransomware	Ransomware expert
Segundox	MPGH	38	540	2	90	19	99	Keylogger	Keylogger expert
Trudo	MPGH	103	339	3	101	33	109	Backdoor	Novice hacker
AliceBob	MPGH	22	234	1	45	10	54	Wi-Fi cracking tool	Game coin black seller
Ymazho	MPGH	537	940	5	143	34	201	Protocol breaker	Game account hacker

of the egonet as shown in Fig. 8. In Table 4, we show the actual values of indicative users, including the three outliers in the plot.

The cross-platform egonet analysis can enrich the profile of each user significantly. For example, if we were just looking at GitHub, we may not have paid attention to *3vilp4wn* and *Evilcry*. Both of these authors are less active on GitHub (small GitHub egonet), but are quite active in the security forums (large forum egonet). A closer investigation of the online forums reveals activities that match their interests on GitHub. This suggests that their GitHub activity is part of their online brand. For example, *3vilp4wn* advertises her GitHub keylogger repository in the forum. Furthermore, she also provides advice on how to develop this type of malware and proposes to form a collaboration group. Similarly, *Evilcry* claims to be a botnet and ransomware expert, especially, for the "WannaCry" ransomware.

One interesting finding from the egonet of author *Ymazho* is that we find his 3 GitHub friends are also active and supporting him in MPGH forums. Together, they form a mini hacking group and they are being contacted by other MPGH users to hack the rivals' gaming account. *Ymazho* is also famous for trading gaming gold coins in black market (acquired by achievements in online games). We intend to expand in this promising direction in the future.

c. Using information from the web In our approach, we leverage existing information on hackers from (a) security outlets and databases and (b) using web queries. With our python-based query and analysis tools, we verified the role and activities of authors, which we omit due to space limitations.

8 Related works

Studying the dynamics of the malware ecosystem on GitHub has received very little attention. Most studies differ from our work in that: (a) they do not focus on malware on GitHub, and (b) when they do, they do not take an author-centric angle as we do here: they focus on classifying malware repositories or use a small set for a particular research study.

Our work builds on our earlier effort (Rokon et al. 2020), whose main goal is to identify malware repositories on GitHub at scale, but it does not study the malware author ecosystem as we do here.

a. Studies of malware repositories on GitHub Several other efforts have manually collected a small number of malware repositories with the purposes of a research study (Lepik et al. 2018; Zhong et al. 2015). Some other studies (Calleja et al. 2016, 2018) analyze malware source code from a software engineering perspective, but use only a small number of GitHub repositories as a reference.

b. Studies of benign repositories on GitHub Many studies analyze benign repositories on GitHub from a point of view of software engineering or as a social network. Some efforts find influential users and analyze the motivation behind following, forking and contributions (Blincoe et al. 2016; Jiang et al. 2017). Earlier efforts study repositories by analyzing the repository-repository relationship graph (Thung et al. 2013) and by using an activity graph (Xavier et al. 2014).

Several works in this area identify influential authors and repositories using: the starring activity (Hu et al. 2016), the following star-fork activity (Hu et al. 2018) or a rank-based approach (Liao et al. 2017). Note that a version of the hyperlink-induced topic search algorithm (Li et al. 2002) has been used by some of the above efforts for calculating influence,

but they do not adjust the weights to account for the different frequencies of the types of interactions between users.

For our bipartite clustering, we adapt the greedy modularity maximization approach (Clauset et al. 2004; Alzahrani and Horadam 2016).

c. Studies on security forums This is a recent and less studied area of research. Most of the works focus on extracting entities of interest in security forums. An interesting study focuses on the dynamics of the black market of hacking goods and services and their pricing (Portnoff et al. 2017). Other studies focus on identifying important events and threats (Sapienza et al. 2017, 2018; Islam et al. 2020b, 2021). None of the aforementioned works focus on the dynamics among hackers across platforms.

d. Cross-platform study Finally, some efforts study author activities on different software development forums, namely GitHub, Stack Overflow and Security Forums (Hauff and Gousios 2015; Lee and Lo 2017; Islam et al. 2020a), but do not consider information from security forums.

9 Conclusion

We develop a systematic approach for studying the ecosystem of hackers. Our approach develops methods to identify (a) influential hackers, (b) communities of collaborating hackers and (c) their cross-platform interactions. Our study concludes in three key takeaway messages: (a) the malware ecosystem is substantial and growing rapidly, (b) it is highly collaborative, and (c) it contains professional malicious hackers.

Our initial findings are just the beginning of a promising future effort that can shed light on this online malware author ecosystem, which spans software repositories and security forums. The current work thus can be seen as a building block that can enable new research directions.

Follow-up research can expand on our work to develop preemptive security initiatives, such as: (a) monitoring hacker activity, (b) detecting emerging trends and (c) identifying particularly influential hackers towards safeguarding the Internet.

Acknowledgements This work was supported by the UC Multicampus-National Lab Collaborative Research and Training (UCNLCRT) award #LFR18548554.

References

Aaron H (2020) 17 years old boy tried to hack twitter. https://bit.ly/3o7zRQl

Alzahrani T, Horadam KJ (2016) Community detection in bipartite networks: algorithms and case studies. In: Complex systems and networks, Springer, pp 25–50

Blincoe K, Sheoran J, Goggins S, Petakovic E, Damian D (2016) Understanding the popular users: following, affiliation influence and leadership on github. Inf Softw Technol 70:30–39

Calleja A, Tapiador J, Caballero J (2016) A look into 30 years of malware development from a software metrics perspective. In: International symposium on research in attacks, intrusions, and defenses, Springer, pp 325–345

Calleja A, Tapiador J, Cabalero J (2018) The malsource dataset: quantifying complexity and code reuse in malware development. IEEE Trans Inf Forensics Secur 14(12):3175–3190

Clauset A, Newman ME, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70(6):6

Cybersec (2018) Stealing password in 5 minutes using wifiphisher. https://www.secjuice.com/phishing-with-wifiphisher/

EmpireProject (2018) Project empire. https://github.com/EmpireProject/Empire

Gharibshah J, Papalexakis EE, Faloutsos M (2020) REST: a thread embedding approach for identifying and classifying user-specified information in security forums. ICWSM

Hauff C, Gousios G (2015) Matching github developer profiles to job advertisements. In: 2015 IEEE/ACM 12th working conference on mining software repositories, IEEE, pp 362–366

Hu Y, Zhang J, Bai X, Yu S, Yang Z (2016) Influence analysis of github repositories. SpringerPlus 5(1):1–19

Hu Y, Wang S, Ren Y, Choo KKR (2018) User influence analysis for github developer social networks. Expert Syst Appl 108:108–118

Islam R, Rokon MOF, Darki A, Faloutsos M (2020a) Hackerscope: The dynamics of a massive hacker online ecosystem. In: 2020 IEEE/ACM International conference on advances in social networks analysis and mining (ASONAM), pp 361–368

Islam R, Rokon MOF, Papalexakis EE, Faloutsos M (2020b) 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), pp 515–522

Islam R, Rokon MOF, Papalexakis EE, Faloutsos M (2021) Recten: a recursive hierarchical low rank tensor factorization method to discover hierarchical patterns in multi-modal data. In: Proceedings of the international AAAI conference on web and social media

Jiang J, Lo D, He J, Xia X, Kochhar PS, Zhang L (2017) Why and how developers fork what from whom in github. Empir Softw Eng 22(1):547–578

Lee RKW, Lo D (2017) Github and stack overflow: analyzing developer interests across multiple social collaborative platforms. In: International conference on social informatics, Springer, pp 245–256

Lepik T, Maennel K, Ernits M, Maennel O (2018) Art and automation of teaching malware reverse engineering. In: International conference on learning and collaboration technologies, Springer, pp 461–472

Li L, Shang Y, Zhang W (2002) Improvement of hits-based algorithms on web documents. In: Proceedings of the 11th international conference on World wide web, pp 527–535

Liao Z, Jin H, Li Y, Zhao B, Wu J, Liu S (2017) Devrank: mining influential developers in github. In: GLOBECOM 2017-2017 IEEE global communications conference, IEEE, pp 1–6

Mitre (2019) State sponsored hacking tool. https://attack.mitre.org/software/S0363

Online Forums (2021) Ethical hacker, hack this site, offensive community, wilders security. https://www.ethicalhacker.net/, https://www.hackthissite.org/, http://offensivecommunity.net/, https://www.wilderssecurity.com/, https://mpgh.net/

Pastrana S, Thomas DR, Hutchings A, Clayton R (2018) Crimebb: Enabling cybercrime research on underground forums at scale. In: WWW, pp 1845–1854

Portnoff RS, Afroz S, Durrett G, Kummerfeld JK, Berg-Kirkpatrick T, McCoy D, Levchenko K, Paxson V (2017) Tools for automated analysis of cybercriminal markets. In: WWW, p 657

Rokon MOF, Islam R, Darki A, Papalexakis EE, Faloutsos M (2020) Sourcefinder: Finding malware source-code from publicly available repositories in github. In: 23rd international symposium on

research in attacks. Intrusions and defenses (RAID), USENIX, pp 149–163
Sapienza A, Bessi A, Damodaran S, Shakarian P, Lerman K, Ferrara E (2017) Early warnings of cyber threats in online discussions. In: 2017 IEEE international conference on data mining workshops (ICDMW), pp 667–674
Sapienza A, Ernala SK, Bessi A, Lerman K, Ferrara E (2018) Discover: Mining online chatter for emerging cyber threats. In: Companion proceedings of the web conference 2018, international world wide web conferences steering committee, WWW '18, pp 983–990
Sophron (2014) Wifiphisher. https://github.com/wifiphisher/wifiphisher. Accessed 14 Mar 2020
Thung F, Bissyande TF, Lo D, Jiang L (2013) Network structure of social coding in github. In: 2013 17th European conference on software maintenance and reengineering, IEEE, pp 323–326
Weng J, Lim EP, Jiang J, He Q (2010) Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on web search and data mining, pp 261–270
Xavier J, Macedo A, de Almeida Maia M (2014) Understanding the popularity of reporters and assignees in the github. In: SEKE
Zhong X, Fu Y, Yu L, Brooks R, Venayagamoorthy GK (2015) Stealthy malware traffic-not as innocent as it looks. In: 2015 10th international conference on malicious and unwanted software, IEEE, pp 110–116