Digital Drift and the Evolution of a Large Cybercrime Forum

Jack Hughes
Department of Computer Science and Technology
University of Cambridge
Cambridge, UK
Jack.Hughes@cl.cam.ac.uk

Alice Hutchings
Department of Computer Science and Technology
University of Cambridge
Cambridge, UK
Alice.Hutchings@cl.cam.ac.uk

***Abstract*—Cybercrime forum datasets are large and complex. Prior research uses aggregated time series data to create a picture of the whole dataset, or focuses on a smaller sample of cross sectional data, often for a specific subcommunity or crime time. This paper uses the longitudinal time series aspect of cybercrime forums to measure and observe the evolution of forums at a macro scale. Applying the digital drift theoretical framework, borrowed from criminology, we find a large amount of churn on the forum, with only a small proportion of users continuing long-term engagement. Measurements show a continual shift in forum activity, with year-based cohorts moving from starting in hacking discussions, towards starting in general discussions, and later towards e-whoring boards. The group of members who are active on the forum for over 12 months, typically have their last post in the marketplace, while other members, who are active for shorter periods of time, have their last post in hacking-related boards. Overall, we see an increasing trend towards financially-driven cybercrime, at both the user and forum level. Users post more in financially-related boards over time, and forum activity has trended away from gaming/social activity, trending towards more activity in market-related boards.**

1. Introduction

Cybercrime forums have been studied in detail, from economic analysis of marketplaces [1], [3], [4], [13]–[15], [23], [28], [31], [39]–[41], to identifying individual key actors [6], [32], [37], to exploring certain types of activities in detail [5], [10], [19], [33], such as e-whoring [19], [33] (fraud where intimate images are used for financial gain). While much of this work has focused on specific aspects, such as analysing cross sectional data at one point in time, we take a macro view of the forum over time, to explore how these forums have evolved. While these forums have a large number of users, only a small fraction of these members continue to remain active, while other members "churn" or "drift". In addition, there is an increasing shift towards marketplace related discussions after joining the forums.

Cybercrime forums are not static. They evolve over time: growing from a small set of users to a large ecosystem of many subcommunities, and sometimes reducing back down in size. Datasets constantly evolve, and need to be kept up to date to follow and track changes. Changes on the forums can be observed at three different levels:

Micro: At the micro layer, this involves observing the changing interests and aims of users over time [6], [7], [32], [37], which can be grouped into pathways on the forum. This can include types of interactions made: moving from social posts to tutorial posts, moving from only posting in existing threads to creating their own, sharing tutorials, participating in wider general social activity, and engaging in marketplace activity. While users may post to the same board about the same topic, there could be change of intention e.g. development of a tool, followed by selling the tool.

Meso: At the meso layer, subcommunities of members discuss topics relevant to their interests, such as to discuss a particular crime type [5], [10], [19], [33], [33]. These subcommunities change over time, growing and shrinking at differing rates. They may use jargon unique to their subcommunity, and contain both members focused solely on a particular type of activity and members spread across different areas of the forums. The definition of these subcommunities depends on what the research task is.

Macro: At the macro level are the large scale forum trends over time [1], [13], [15], [28], [39], including across forums, such as measuring the rising and falling of keywords and boards, across all members. Also, tools such as a dictionary/tool for detecting jargon in the forum (collecting spelling variations and matching to general categories) can be used to track changes. Other large scale trends include the churn of users on the forum: those that join and leave in the same year may have different intentions to those that stay for multiple years.

Most macro research of forums uses the data as a "static" dataset to collect aggregate statistics over the entire forum. While aggregate statistics are useful for measurements, users are not stable data points. Topics of interest and activity levels vary over time, and longitudinal analysis is needed to understand how this engagement fluctuates, which is the main contribution of this paper.

The work uses data from CrimeBB, a dataset available for researcher use from the Cambridge Cybercrime Centre, consisting of over 20 forums and 100 million posts spanning over 20 years. Focusing on HackForums, the largest and longest-running English-language forum, we measure board activity, churn of users on the forum who drift off the platform, and shifting interests over time for users with different levels of activity. This paper explores how different cohorts have varied in scale over time, and changes in different topic interests. We show how users churn on the forum, with year-based cohorts

halving in size every 12 months. This paper takes a longitudinal approach at analysing and describing changes to this dataset over time, which is useful for later research into cybercrime pathways, as the topics of interest and intentions of users changes over time.

We use the digital drift theoretical framework to explore trends in user cohorts, and the forum overall, over time. This theory prompts us to think about the continual ebbs and flows of user engagement. As new members join and other members leave, we hypothesise a change in topics on the forum due to the large changing community. For example, a topic may be popular in one year and then decrease as users drift off the platform and new users join. Also, as the community changes, the value of the community to members changes, leading to fluctuations in retention levels. Drift theory helps to us to explore ideas around the changing community and the effect this has on the forum. We address the following research questions:

- How has the size of the forum, and topics of discussion, changed over time?
- What are popular topics among user cohorts over time?
- How do retention levels change over time?
- With new members joining the forum and old members leaving, what effect does this have on the interests of forum members?

We study the effect on the forum of churning, and use topic models explain how the cohort interests have changed over time. We find that over time, the forum has become more financially driven, in terms of the activity on the forum as a whole and as groups of users turn to the economic aspects over time.

Our contributions include:

- Categorisation of members to filter active users from inactive users, to separate analysis of users that post a few times and leave from those that remain active on the forum for multiple months
- Measurements of churn on the forum, finding only a small proportion of users continue engagement
- Measurements of starting on the forum, including finding a shift from beginner hacking discussions towards general, then towards e-whoring discussions. This is also shown to occur month-on-month after joining the forum.
- Measurements of last posts on the forum show for members posting less than 12 months, they post on hacking boards, whereas members who are more active are typically posting in the marketplace before leaving

2. Background

Cybercrime forums vary in type. Some can be found on the open ('surface') web, and others on the dark web, which operate as hidden services on anonymity networks. Larger forums cater to English speakers, although other-language forums exist [2], [17]. Forums can also specialise on certain topics, e.g. carding (credit card fraud) [26], [38], or be more general, with a wide range of different boards (high-level categories of topics) available [32]. The size and topic of a forum can affect the social network structure the forum [36]. Also, some forums have a reputation system, which can be used as a proxy to trust other members within the forum [13], [32].

2.1. Social Network Analysis & Churn

This work combines perspectives across many fields, including theory from criminology. The social aspects of cybercriminals has been studied across criminological literature and quantitative social science [11], [24], [27], [30], including the application of social network analysis approaches [12], [29], [35], ranging from studying small real-world criminal groups to larger online communities. This has included using ground truth data following arrests, to build up real-world social networks of key hackers and model how a network could be destabilised [12], and the use of time-series data in longitudinal network analysis for exploring online ecosystems of hacking defacement groups [35].

One approach for observing the evolution of forums is to measure how the social network structure has changed over time. Public data from cybercrime forums typically does not include direct user to user interactions, such as a 'friend' mechanism. Instead, the social graph has to be approximated, with some limitations. For example, a user to user connection could exist by users replying in the same thread, or a user to topic or user to board graph could be built to show interests of the forum. Prior work has used the thread-reply approach to approximate the social graph [36], which found some highly connected members who form hubs, posting general discussions and tutorials on broader topics, and most members have a lower number of connections. This approximation is limiting for measurements, as is can cause a bias in results towards those who create threads. Using this approach, the thread creator acts as the 'hub' within the social network, inflating their importance within the network without regard to the quality of the thread.

This work explores the churn of users. Churn can be defined as users leaving the forum and not returning. Churn in general social networks is commonly used for prediction tasks that use feature-based approaches [20]. The aim of this task is to identify which users will churn, and to calculate the churn rate (number of members leaving in a period divided by active members). For example, if "popular" members churn, then other members may be likely to churn. Churn analysis has commonly been used in industry to identify "high-value" users who may leave, so companies can target retention efforts. However, in industry, such as with telecoms, users enter a service contract with a switching cost. Forums contain a much lower switching cost, and members are weakly-tied. Within forums, if the number of users becomes too small, and network effects diminish, this can cause a decrease in marketplace activity and forum posts. There are differing incentives at play here. Forum administrators will want to retain users so their community continues to flourish, while still actively moderating their platform, removing members who break platform rules. Counter to forum administrators wanting to avoid the death of a forum, interventions by law enforcement aim to disrupt cybercrime economies, with the removal of key actors

likely to have a disproportionately large effect on the network.

There is not one single type of churn [21]. Churn can be typical (members stop posting), holiday (members stop posting for a period of time but later return), bursty behaviour (members post infrequently, leading to misidentification of churn), and inconsistent behaviour. Churn has been used to explore the relation between a user's value in a community and the probability of a user churning [22]. This used a user-to-user thread reply graph for features, with edges weighted by the number of replies, and forum specific metrics for users. Metrics include the average length of posts started by a user, average length of thread the user participates in, popularity of posts, initialisation of posts, and polarity.

Less work has focused specifically on churn within cybercrime forums. One study used private message interactions from three carding forums [15]. They compare the change in topology (structure) of the network between members banned on the forum to the regular churn of users on the forum. While two of the forums did not have significant results due to a low number of bans, one forum found the change in small world structure between the two types was similar, finding that bans did not affect the overall topology. The authors also use the Louvain method to cluster communities in the forums, with latent Dirichlet allocation (LDA) topic models to discover topics, to identify specialities. They find most communities sub-specialise in specific crimes, with smaller two-tiered communities of 100-230 members, and larger multi-tiered communities. While the study is useful, it is not clear if this generalises to larger general-purpose hacking and cybercrime forums.

2.2. Digital Drift

Similar concepts to social network churn exist in criminology, with the 'drift' [25] and 'digital drift' [16], [18] frameworks, which explain how drift into and out of criminal pathways can often be 'accidental' or 'unpredictable' [18]. In drift theory, Matza points out offenders drift in and out of crime, enabled by a loosening of social control [25]. According to digital drift [16], [18], an application of drift theory, forums and platforms provide a mechanism for engaging and disengaging from discussions about hacking and crime. In this study, we focus more on measuring the drift aspects of the framework, rather than the social control elements of the theory. The combination of uniting digital drift theory with the practical aspects of social network measurements provides a unique perspective in the field. We use measurements from standard longitudinal approaches and topic modelling, instead of using social network graph approaches, as we do not have ground truth data of social interactions, and approximating these can lead to inaccurate results.

3. Methods

3.1. Data

This work uses data from CrimeBB [34], available from the Cambridge Cybercrime Centre for academic research use under a data sharing agreement. This has data for over 20 years on underground hacking forums, with differing language, size, and topics. The majority of this paper focuses on the HackForums subset of CrimeBB forum which is the largest forum in the dataset. This dataset is selected as the timescale of data available allows for longitudinal analysis of the evolution of the forum, and the scale provides a more representative view of the open cybercrime platform ecosystem. However, we note that findings in this paper may not reflect activity patterns found on smaller closed and more profit-oriented cybercrime discussion platforms. This contains over 680,000 users, 42,000,000 posts, 4,000,000 threads, from 2007 to 2020. This subset also includes reputation scores: positive or negative votes can be sent between users to build a feedback system, as a proxy for trust. While the forum is quite large in size, the majority of members are inactive lurkers, drifting off posting activity. As the dataset only contains posts, not views, we are unable to tell if these members have completely stopped engaging with the forum. There is a small concentration of highly active users, and a subset of these may be considered 'key actors' in the cybercrime literature: members of the forum that could be of interest to law enforcement.

As the dataset comes from an automated scraper, there is likely to be occasional data quality issues, and as the data is unprocessed, data cleaning needs to be carried out. For topic modelling work, we pre-tokenise the dataset using nltk's [8] TweetTokenizer, and store this tokenised form in the database for quick retrieval. Thread creation dates are not available, but this can be approximated from the first post creation date in the thread. First posts can be obtained by sorting by post id: to overcome an issue where the dataset had incorrectly parsed AM and PM timestamps, and therefore sorting by date for a thread that has a conversation over midnight can cause the ordering to be incorrect. In some cases, where a member has permanently deleted their account, the database will use the value '-1' as a placeholder user ID. Depending on the measurement, this may need to be removed to avoid results containing a very active '-1' user. With all measurements, it's important to lookout for anomalies in results which could be caused by data quality issues or limitations.

3.2. Categories of Users By Level of Activity

Different forums have varying numbers of users. However, as these are types of social networks, they tend to have a 'long tail' of activity: few users post the majority of the content. Figure 1 breaks this down into categories of activity levels, and Table 1 lists the number of members per category. The -1 category is for posts which the user has since deleted their account, with the placeholder ID '-1'. The second category (Cat 0) is of users that have less than 12 months of activity on the forum. Those with 12 or more months of activity are broken down into thirds of posting volume. The categorisation is useful to visualise the level of activity of forum members, and to begin to sample from the forum. Sampling is essential, as it is not trivial to get meaningful findings from the forum as a whole due to the size of these.

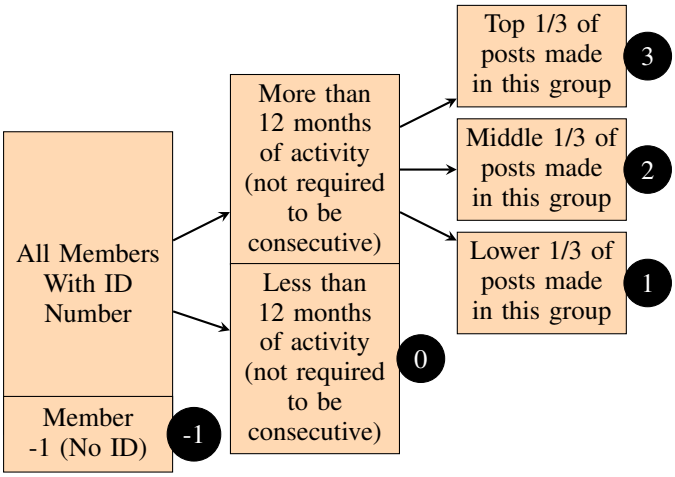


Figure 1: Categories on the forum by level of activity

3.3. Analytical Approach

In this work, we use the categories of activity levels to explore effects of drifting and churn on the forum, how the interests of users have changed over time, and their first and last posts. We use latent Dirichlet allocation (LDA) topic models to observe the trends of moving between first and last topics on the forum. We use coherence scores to select the number of topics with the greatest similarity for documents (posts) per topic. We combine topics for measurement where they have a strong overlap (e.g. "general" topics).

We measure drift and churn by looking at activity levels declining over time per type of board as defined by Pastrana et al. [32], namely common, hack, tech, coding, games, market, money, web, and graphics. We also analyse year-on-year cohorts, to identify proportions of users that remain active on the forum beyond the year they join in.

We measure the interests of years using topics and boards. As the forum userbase grows, the forum takes on a wider range of interests and the topics included start to change over time. We also measure the proportions of users replying to a new thread (typically contributing something new or asking a question) against those replying to an existing thread.

3.4. Ethics

Ethics approval was granted from the department's ethics committee for this research. We used data collected from a publicly available forum, and could not gain informed consent from all members as this would be considered to be spamming. As we only analyse posts as a collective whole (i.e. aggregated into time periods or large groups), rather than identifying individual users, under the British Society of Criminology's Statement of Ethics [9], this falls outside of the requirement of informed consent. We also avoid publishing details that could identify individuals, including usernames and original post contents.

TABLE 1: Table of categories on the forum by level of activity

Category	Number of Members
-1	Unknown
0	601,642
1	12,789
2	12,815
3	13,211

3.5. Limitations

Our methods and dataset are subject to limitations. Members who have deleted their accounts prior to data collection will not have their data available. A placeholder value (-1) is used instead, and it is not possible to match posts together. Therefore, we are unable to measure start and end posts for this group. We exclude this group from our analysis, but note results may be limited due to this.

The dataset only contains data for actively posting users. It does not include users that register with the forum and read about topics, consuming material but never actually posting publicly. This can result in bias.

The dataset has post data for each forum profile, and we note that forums users may use more than one profile. Stylometric methods could be used to group these profiles together, however our analysis assumes each profile is a unique user and we leave this further analysis to future work.

Members who post a single time on the forum and leave their accounts idle will be grouped into Category 0. We focus our later analysis on members with regular posting activity to explore changes over time, and note that our analysis does not explore deeply this group of single post accounts.

Forum data contains a partial structure. Board categories, boards, and threads provide 'ground-truth' structure to measure, but post content does not. Topic models can provide an approximation of discussions, but we note this is not a perfect representation of content. We choose parameters using coherence scores and manually validate detected topics, but the number of topics chosen is selected by the user (i.e. there is not a "perfect" number of topics). A greater number of topics can be chosen to create a granular model, however this can be harder to visualise.

Measurements show a shift towards marketplace activity from general chatter and hacking discussions. It is important to note that these are found in aggregated data: while there is an overall shift, members individually move into marketplace discussions at different rates and will each be interested in different topics. Overall trends are representative of groups, not individual users.

4. Results

We split our measurements into three themes. First, activity over the entire timespan of the forum to explore churn on the forum. Second, exploring how users get started on the forum, including which boards members post to first and continue to engage in for up to 6 months. Third, we explore declining activity levels for the users who had shown the most commitment to the forum (more than 12 months of activity).

4.1. Activity From Joining to Leaving

It is readily apparent that there is a large amount of drifting and churn occurring on the forum. When users engage with posts on boards for the first time, some users may continue to sustain an interest in these boards, while a significant proportion 'drift' away. Along with drifting, the forum also experiences churn. Instead of users changing their interests and moving away from boards, 'churn' looks at how long year-based cohorts remain on the forum. Across all activity categories, Figure 2 shows that only smaller groups of each year cohort continue to interact with posts and threads on the forum over time. We see forum activity peaks in 2012, with a gradual decline thereafter.

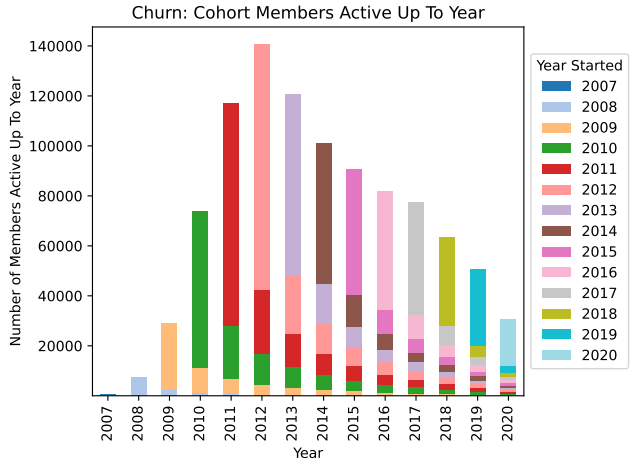


Figure 2: Churn of users on the forum ❶ ❷ ❸ ❹

Note that a small cohort of users who joined each year continue to be active in all subsequent years, including those who joined in 2007. This is more visible in Figure 3, which visualises the proportion of active users in each year after joining the forum. We can see the year 2009 has the highest retention of users, which decreases year on year.

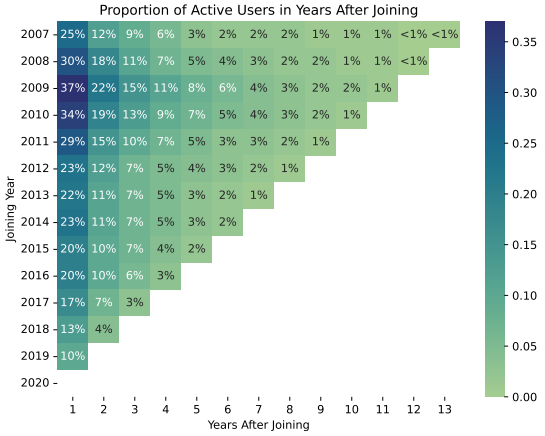


Figure 3: Active Users Over Time Per Joining Year ❶ ❷ ❸ ❹

We further explore the first and last posts of forum members, starting with the type of first post made, and measuring the change in topics from first to last post. Measuring whether a first post was a new thread or a reply in an existing thread is useful to show the difference between new users who first post to existing threads, contributing to existing discourse, compared to users opening new threads to request or share information and tools. We observed that the number of new threads remains steady over time, but the number of replies increases sharply with an influx of new users between 2010 and 2013. A majority of users reply to existing threads for their first post on the forum.

Over the entire duration of activity on the forum, users may change topics they are interested in. We train two latent Dirichlet allocation (LDA) topic models over first and last posts on the forum, using coherence scores to select the number of topics that each have the highest similiarity within. We then validate the topics, to check they represent real concepts.

Using these two models, we plot a Sankey diagram of first post topic to last post topic, shown in Figure 4. This shows 'general' and 'requests' are the most popular topics to post about. Other joining topics include discussions of online accounts, remote access trojans, and packs. Leaving topics include general help, general discussions of accounts, money, and marketplace. General discussions may include user introductions, with topic words including 'I'm new', the other topics provide clues into some of the reasons why users chose to join the forum.

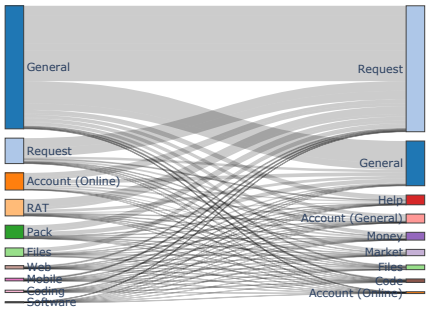


Figure 4: First Post Topic to Last Post Topic ❶ ❷ ❸ ❹

4.2. Starting Off

In this subsection, we explore the start of topic and activity pathways that exist on the forum. By breaking the first posts down by board per year, we can observe where these posts concentrate over time. Figure 5 shows the first post as a proportion of max in a year (max is 1.0, other boards are a proportion of this). This shows the top 10 boards, rather than all. Using the year-by-year proportional heatmap, we observe that new members started posting in Beginner Hacking up to 2012. When the new user counts were at a peak, new members made their

first post in Hacking Tools and Programs. Then, following a decline in the number of new users, The Lounge and Introductions were used to make their first post. More recently, since 2019 e-whoring, in which users use 'packs' of explicit images to take money from victims, has been a popular board for users making their first post.

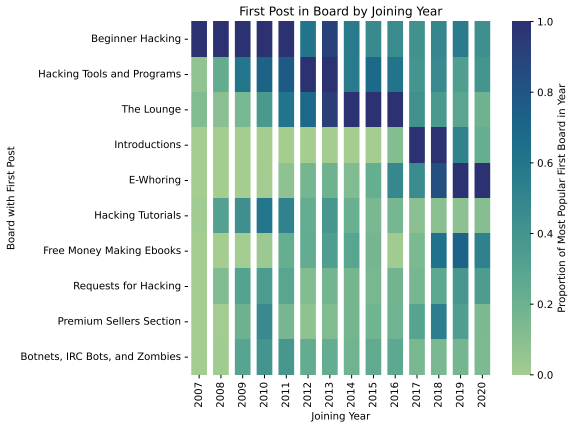


Figure 5: First Post in Boards (Top 10) ⓪ ① ② ③

While Figure 7 shows how users' first posts change over time, we can also observe how the cohorts evolve, with changes in the boards posted to by members joining in various years. Figures 6, 7, and 8 show members posting to beginner hacking for the first time in 2010 and 2015 and 2019 respectively, with a column for each month after the joining month. We select 2010 as this is when forum has become established, 2015 follows peak activity on the forum, and 2019 is the last full year of data collected. We can see that the interests of year cohorts changes over time. Colour coding is used to indicate the type of board: 'Financial' for financial-related discussions including market and money making activity, 'Gaming/Social' for both gaming and general social chat, and 'Knowledge' for exchange of ideas and tools.

The 2010 cohort (Figure 6) has most users engaged in knowledge boards around hacking techniques, with a smaller proportion interested in general chat. Over time, we observe a shift to the majority of members involved in general chatter and marketplace related boards. The 2015 cohort (Figure 7) is immediately more interested in general chatter, and there also are discussions of gaming. Again, there is a shift towards marketplace related discussions beyond the first month.

The 2019 cohort (Figure 8) is primarily interested in financial related boards, specifically 'e-whoring', and general chatter. Overall, we see a marked reduction in knowledge exchange boards, and an increase in interest in money making and market-related boards. This includes both at the cohort level (within each figure) and across the forum as a whole. Gaming and social related boards remain important to users throughout, indicating that there is a consistent level of community building and social interaction that is important over time.

4.3. Declining Activity

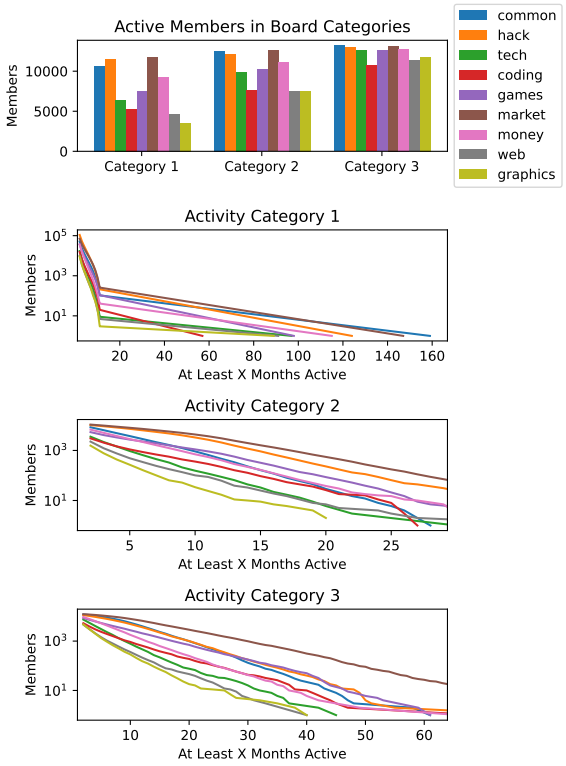


Figure 9: Members posting to categories over time ① ② ③

The digital drift theoretical framework hypothesises that members move away from activities, often unpredictably. We use this framework to explore how activity declines over time for different activity categories, comparing the types of boards posted to by those with low, medium, and high post counts. Figure 9 shows drifting occurring on various categories within activity categories ① ② ③

The first subplot highlights the difference between activity categories. We can see that the most active users show less variability in the types of boards they post in, while those that post less often are more likely to post in the market and hack categories.

The following subplots highlights the declining activity of members over multiple months within each group. The market category contains the highest level of activity over time, and hacking boards contain a sharper decline over time compared to these.

While these subplots highlight differences between activity category groups, they assume that all members start on the same month. We next take an alternative approach, shown in Figure 10. This shows the board in which users posted their last post, by year, aggregated into a small set of categories. This aggregation adapts Pastrana et al. [32] who categorised boards into the following categories: Coding, Common, Games, Graphics, Hack, Market, Money, Tech, and Web. We add boards established since this earlier work to the existing categorisation. For activity category ⓪ across all years 'hack' is the most popular board type, with a gradual rise in popularity of

Proportion of Members Active in Each Board (Each Month After Joining) - Starting in 2010 with First Post in Any Board

0 Months After Joining	1 Month After Joining	2 Months After Joining	3 Months After Joining	4 Months After Joining	5 Months After Joining	6 Months After Joining
Beginner Hacking 20.9%	The Lounge 7.71%	The Lounge 5.92%	The Lounge 5.21%	The Lounge 4.87%	The Lounge 4.76%	The Lounge 4.62%
Hacking Tools and Programs 16.7%	Beginner Hacking 6.88%	Premium Sellers Section 4.86%	Premium Sellers Section 4.08%	Buyers Bay 3.64%	Buyers Bay 3.38%	Buyers Bay 3.19%
Hacking Tutorials 14.98%	Hacking Tools and Programs 6.83%	Buyers Bay 4.59%	Buyers Bay 3.96%	Premium Sellers Section 3.6%	Premium Sellers Section 3.22%	Premium Sellers Section 3.06%
The Lounge 13.73%	Premium Sellers Section 6.73%	Hacking Tools and Programs 4.41%	Hacking Tools and Programs 3.69%	Hacking Tools and Programs 3.08%	Secondary Sellers Market 2.72%	Secondary Sellers Market 2.84%
Premium Sellers Section 11.85%	Buyers Bay 6.16%	Beginner Hacking 4.33%	Beginner Hacking 3.35%	Beginner Hacking 2.84%	Hacking Tools and Programs 2.68%	Hacking Tools and Programs 2.57%
Buyers Bay 10.75%	Hacking Tutorials 5.37%	Hacking Tutorials 3.56%	Hacking Tutorials 2.89%	Secondary Sellers Market 2.7%	Beginner Hacking 2.63%	Free Services and Giveaways 2.37%
Remote Administration Tools 9.17%	Remote Administration Tools 4.91%	Referrals 3.33%	Referrals 2.79%	Referrals 2.44%	Free Services and Giveaways 2.26%	Beginner Hacking 2.34%
Referrals 8.44%	Referrals 4.51%	Remote Administration Tools 3.15%	Service Offerings 2.71%	Hacking Tutorials 2.43%	Referrals 2.19%	Referrals 2.21%
Keyloggers 7.16%	Service Offerings 4.14%	Service Offerings 3.05%	Remote Administration Tools 2.61%	Service Offerings 2.42%	Monetizing Techniques 2.18%	Rules, Announcements, News, and Feedback 2.17%
Requests for Hacking 7.01%	Monetizing Techniques 3.68%	Monetizing Techniques 2.76%	Secondary Sellers Market 2.58%	Monetizing Techniques 2.27%	Service Offerings 2.18%	Monetizing Techniques 2.01%

Financial
Gaming/Social
Knowledge

Figure 6: Members posting to any board for the first time in 2010 ❶ ❷ ❸

Proportion of Members Active in Each Board (Each Month After Joining) - Starting in 2015 with First Post in Any Board

0 Months After Joining	1 Month After Joining	2 Months After Joining	3 Months After Joining	4 Months After Joining	5 Months After Joining	6 Months After Joining
The Lounge 17.43%	The Lounge 5.96%	The Lounge 3.34%	The Lounge 2.64%	The Lounge 2.29%	The Lounge 1.92%	The Lounge 1.83%
Hacking Tools and Programs 12.97%	Marketplace Discussions 3.27%	Premium Sellers Section 1.97%	Premium Sellers Section 1.7%	Premium Sellers Section 1.44%	Premium Sellers Section 1.31%	Premium Sellers Section 1.3%
Beginner Hacking 11.11%	Premium Sellers Section 3.12%	Marketplace Discussions 1.79%	Marketplace Discussions 1.4%	Marketplace Discussions 1.2%	Buyers Bay 1.09%	Buyers Bay 1.02%
Grand Theft Auto 7.04%	Beginner Hacking 3.09%	Buyers Bay 1.62%	Buyers Bay 1.33%	Buyers Bay 1.12%	Marketplace Discussions 0.99%	Free Services and Giveaways 0.96%
Marketplace Discussions 6.8%	Hacking Tools and Programs 3.06%	Beginner Hacking 1.49%	Counter Strike 1.22%	Free Services and Giveaways 1.09%	Counter Strike 0.98%	Secondary Sellers Market 0.92%
Premium Sellers Section 6.35%	Buyers Bay 2.34%	Counter Strike 1.48%	Free Services and Giveaways 1.19%	Counter Strike 1.08%	Free Services and Giveaways 0.97%	Marketplace Discussions 0.87%
E-Whoring 6.02%	Free Services and Giveaways 2.33%	Hacking Tools and Programs 1.47%	Beginner Hacking 1.13%	Secondary Sellers Market 0.99%	Secondary Sellers Market 0.88%	Counter Strike 0.86%
Counter Strike 5.43%	Counter Strike 2.23%	Free Services and Giveaways 1.44%	Hacking Tools and Programs 1.06%	E-Whoring 0.96%	Beginner Hacking 0.81%	E-Whoring 0.79%
Remote Administration Tools 4.89%	E-Whoring 2.22%	E-Whoring 1.27%	Secondary Sellers Market 1.06%	Beginner Hacking 0.96%	E-Whoring 0.78%	Beginner Hacking 0.7%
Botnets, IRC Bots, and Zombies 4.87%	Remote Administration Tools 2.03%	Secondary Sellers Market 1.26%	E-Whoring 1.02%	Hacking Tools and Programs 0.82%	Hacking Tools and Programs 0.68%	Hacking Tools and Programs 0.62%

Financial
Gaming/Social
Knowledge

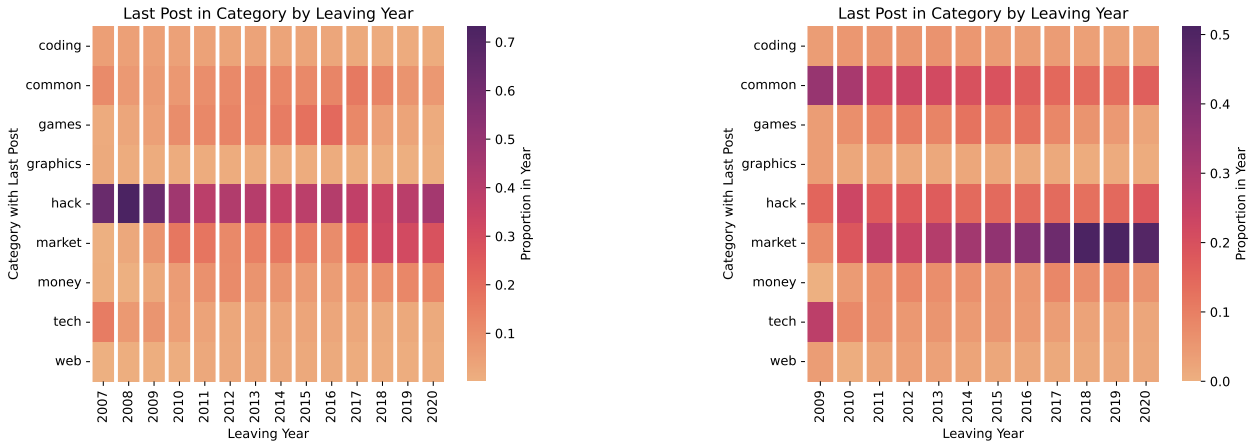
Figure 7: Members posting to any board for the first time in 2015 ❶ ❷ ❸

Proportion of Members Active in Each Board (Each Month After Joining) - Starting in 2019 with First Post in Any Board

0 Months After Joining	1 Month After Joining	2 Months After Joining	3 Months After Joining	4 Months After Joining	5 Months After Joining	6 Months After Joining
E-Whoring 12.5%	E-Whoring 2.24%	The Lounge 1.11%	The Lounge 0.86%	E-Whoring 0.75%	E-Whoring 0.67%	The Lounge 0.57%
Free Money Making Ebooks 10.49%	The Lounge 2.19%	E-Whoring 1.08%	E-Whoring 0.84%	The Lounge 0.74%	The Lounge 0.63%	E-Whoring 0.56%
Beginner Hacking 8.25%	Premium Sellers Section 2.09%	Buyers Bay 1.04%	Premium Sellers Section 0.79%	Premium Sellers Section 0.65%	Premium Sellers Section 0.53%	Buyers Bay 0.39%
The Lounge 6.84%	Buyers Bay 1.9%	Premium Sellers Section 1.03%	Buyers Bay 0.7%	Buyers Bay 0.6%	Buyers Bay 0.48%	Premium Sellers Section 0.39%
Buyers Bay 6.77%	Secondary Sellers Market 1.72%	Secondary Sellers Market 0.84%	Secondary Sellers Market 0.56%	Free Services and Giveaways 0.43%	Secondary Sellers Market 0.41%	Free Services and Giveaways 0.32%
Free Services and Giveaways 6.75%	Free Money Making Ebooks 1.64%	Free Money Making Ebooks 0.72%	Free Services and Giveaways 0.49%	Secondary Sellers Market 0.41%	Free Services and Giveaways 0.36%	Secondary Sellers Market 0.28%
Introductions 6.5%	Free Services and Giveaways 1.55%	Free Services and Giveaways 0.7%	Free Money Making Ebooks 0.48%	Currency Exchange 0.4%	Currency Exchange 0.32%	Free Money Making Ebooks 0.27%
Premium Sellers Section 6.16%	Beginner Hacking 1.29%	Beginner Hacking 0.53%	Beginner Hacking 0.41%	Free Money Making Ebooks 0.38%	Free Money Making Ebooks 0.31%	Currency Exchange 0.24%
Secondary Sellers Market 5.81%	Remote Administration Tools 0.98%	Currency Exchange 0.49%	Currency Exchange 0.39%	Deal Disputes 0.33%	Deal Disputes 0.27%	Deal Disputes 0.23%
Requests for Hacking 5.36%	Marketplace Discussions 0.92%	Remote Administration Tools 0.48%	Deal Disputes 0.37%	Remote Administration Tools 0.31%	Member Contests 0.25%	Member Contests 0.21%

Financial
Gaming/Social
Knowledge

Figure 8: Members posting to any board for the first time in 2019 ❶ ❷ ❸



(a) Activity Category: ❶

(b) Activity Categories: ❶ ❷ ❸

Figure 10: Last Post in Board Category

'market' boards in recent years. For activity categories ❶ ❷ ❸ 'common', 'hack' and 'tech' were popular leaving board categories, but from 2011 onwards, 'market' has become the most popular board type posted to before leaving. Overall, this shows a shift towards posting in the marketplace before leaving, with a more significant shift occurring among members of 12 or more non-contiguous months of activity.

5. Discussion & Future Work

This work has shown different aspects to forum activities, guiding analysis using the digital drift theoretical framework. We observed a significant amount of drift and churn. Only a small proportion of users actively contribute to the forum for over 1 year of activity, with these members contributing to a larger number of unique boards. We also observe a skew of members posting replies to other threads rather than creating their own. The majority of members 'churn', with less than 100,000 active members remaining on the forum over time. Observing the churn year-on-year, we note a large number of users join the forum and leave within a single year, with typically less than half of users remaining. This significant level of churn supports drift theory, with groups of users leaving the forum and a smaller group of more persistent users who remain on the forum for longer.

In measuring the interests of members joining the forum, we first observe movements of those posting in beginner hacking moving to more general and market oriented parts of the forum. Breaking this down further by first post per joining year, we identify a shift from beginner hacking to general boards to the e-whoring board. However, the raw number of members joining is significantly higher during earlier years of the forum. These findings support the digital drift theory, showing movement out of hacking-related posts, towards more general chatter, followed by an emergence of interest in a specific activity type.

Grouping boards into categories shows first posts are typically in the hack category, and last posts are either in the hack or market category. In selecting users that are active for more than a year, a last post in the market category is more likely to occur.

Future work should explore trends in other similar forums. This work was limited to a single forum dataset, which may not be fully representative of the full cybercrime platform ecosystem. However, this forum was selected as it one of the largest cybercrime forums.

Future work should further use topic models to explore how the forum has changed over time, both overall and per year cohorts. Training topic models for large number of documents (posts) for different parameters (number of topics) is computationally expensive. However, it would be useful to explore the set of posts made by members when joining and leaving, and additionally sampling at a suitable frequency in between. This can help to build a set of pathways for groups moving through the forum. These pathways could be used to measure if groups are joining the forum for specific intentions (e.g. hacking), for looking at current popular discussion topics (e.g. e-whoring), and if groups joining for specific intentions drift into the group interested in popular discussion topics.

6. Conclusion

In this paper, we use the digital drift theoretical framework to present an analysis of the evolution of forums over time. This includes a categorisation of members into groups, to filter out active users from inactive users. We measure churn on the forum, finding only a small proportion of users continue engagement over multiple years. Over time, first posts have shifted from hacking boards, to general boards, then to marketplace boards over different years. Within each year, there is a clear shift towards more members becoming active in marketplace related activity. We measure the topics of members leaving the forum, finding those with a lower level of activity have their final post on hacking-related boards, whereas members with over 12 months of activity typically place their last post within the marketplace. Our findings overall show a shift towards financial related activity, but a reduction in forum activity as a whole, supporting the digital drift theory.

Acknowledgements

This work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 949127).

As part of the open-report model followed by the Workshop on Attackers & CyberCrime Operations (WACCO), all the reviews for this paper are publicly available at https://github.com/wacco-workshop/WACCO/tree/main/WACCO-2023.

References

[1] Sadia Afroz, Vaibhav Garg, Damon McCoy, and Rachel Greenstadt. Honor among thieves: A common's analysis of cybercrime economies. In *Proceedings of the APWG eCrime Researchers Summit*, 2013.

[2] Sadia Afroz, Aylin Caliskan Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. Doppelgänger Finder: Taking stylometry to the underground. In *2014 IEEE Symposium on Security and Privacy*, pages 212–226, 2014.

[3] Ugur Akyazi, Michel van Eeten, and Carlos H Gañán. Measuring cybercrime as a service (caas) offerings in a cybercrime forum. In *Workshop on the Economics of Information Security*, 2021.

[4] Luca Allodi. Economic factors of vulnerability trade and exploitation. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1483–1499, 2017.

[5] Gilberto Atondo Siu, Ben Collier, and Alice Hutchings. Follow the money: The relationship between currency exchange and illicit behaviour in an underground forum. In *2021 IEEE European Symposium on Security and Privacy Workshops*, pages 191–201, 2021.

[6] Victor Benjamin and Hsinchun Chen. Securing cyberspace: Identifying key actors in hacker communities. In *IEEE International Conference on Intelligence and Security Informatics*, pages 24–29, 2012.

[7] Rasika Bhalerao, Maxwell Aliapoulios, Ilia Shumailov, Sadia Afroz, and Damon McCoy. Mapping the underground: Supervised discovery of cybercrime supply chains. In *APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–16, 2019.

[8] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition, 2009.

[9] British Society of Criminology. Statement of ethics. http://www.britsoccrim.org/ethics/, 2015.

[10] Andrew Caines, Sergio Pastrana, Alice Hutchings, and Paula J Buttery. Automatically identifying the function and intent of posts in underground forums. *Crime Science*, 7(1):1–14, 2018.

[11] Benoît Dupont and Jonathan Lusthaus. Countering distrust in illicit online networks: The dispute resolution strategies of cybercriminals. *Social Science Computer Review*, 40(4):892–913, 2022. PMID: 35971389.

[12] David Décary-Hétu and Benoit Dupont. The social network of hackers. *Global Crime*, 13(3):160–175, 2012.

[13] Jason Franklin, Adrian Perrig, Vern Paxson, and Stefan Savage. An inquiry into the nature and causes of the wealth of internet miscreants. In *Proceedings of the 14th ACM Conference on Computer and Communications Security*, CCS '07, page 375–388, New York, NY, USA, 2007. Association for Computing Machinery.

[14] Tianjun Fu, Ahmed Abbasi, and Hsinchun Chen. A focused crawler for dark web forums. *Journal of the American Society for Information Science and Technology*, 61(6):1213–1231, 2010.

[15] Vaibhav Garg, Sadia Afroz, Rebekah Overdorf, and Rachel Greenstadt. Computer-supported cooperative crime. In Rainer Böhme and Tatsuaki Okamoto, editors, *Financial Cryptography and Data Security*, pages 32–43, Berlin, Heidelberg, 2015. Springer Berlin Heidelberg.

[16] Andrew Goldsmith and Russell Brewer. Digital drift and the criminal interaction order. *Theoretical Criminology*, 19(1):112–130, 2015.

[17] John Grisham, Sagar Samtani, Mark Patton, and Hsinchun Chen. Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence. In *IEEE International Conference on Intelligence and Security Informatics*, 2017.

[18] Thomas J. Holt, Russell Brewer, and Andrew Goldsmith. Digital drift and the "sense of injustice": Counter-productive policing of youth cybercrime. *Deviant Behavior*, 40(9):1144–1156, 2019.

[19] Alice Hutchings and Sergio Pastrana. Understanding eWhoring. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 201–214, 2019.

[20] Marcel Karnstedt, Tara Hennessy, Jeffrey Chan, Partha Basuchowdhuri, Conor Hayes, and Thorsten Strufe. *Churn in Social Networks*, pages 185–220. Springer US, Boston, MA, 2010.

[21] Marcel Karnstedt, Tara Hennessy, Jeffrey Chan, and Conor Hayes. Churn in social networks: A discussion boards case study. In *2010 IEEE Second International Conference on Social Computing*, pages 233–240, 2010.

[22] Marcel Karnstedt, Matthew Rowe, Jeffrey Chan, Harith Alani, and Conor Hayes. The effect of user features on churn in social networks. In *Proceedings of the 3rd International Web Science Conference*, WebSci '11, New York, NY, USA, 2011. Association for Computing Machinery.

[23] M Kiran Kumar and Dr K Bhargavi. An effective study on data science approach to cybercrime underground economy data. *Journal of Engineering, Computing and Architecture*, 10(1):148–158, 2020.

[24] David Maimon, Andrew Fukuda, Steve Hinton, Olga Babko-Malaya, and Rebecca Cathey. On the relevance of social media platforms in predicting the volume and patterns of web defacement attacks. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4668–4673, 2017.

[25] David Matza. *Delinquency and Drift*. John Wiley & Sons, 1964.

[26] Alexander Mikhaylov and Richard Frank. Cards, money and two hacking forums: An analysis of online money laundering schemes. In *Proceedings of the European Intelligence and Security Informatics Conference*, 2016.

[27] Carlo Morselli. Assessing vulnerable and strategic positions in a criminal network. *Journal of Contemporary Criminal Justice*, 26(4):382–392, 2010.

[28] Marti Motoyama, Damon McCoy, Kirill Levchenko, Stefan Savage, and Geoffrey M Voelker. An analysis of underground forums. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, pages 71–80, 2011.

[29] Marie Ouellet, Martin Bouchard, and Yanick Charette. One gang dies, another gains? the network dynamics of criminal group persistence*. *Criminology*, 57(1):5–33, 2019.

[30] Masarah Paquet-Clouston, David Décary-Hétu, and Olivier Bilodeau. Cybercrime is whose responsibility? a case study of an online behaviour system in crime. *Global Crime*, 19(1):1–21, 2018.

[31] Andrew J. Park, Richard Frank, Alexander Mikhaylov, and Myf Thomson. Hackers hedging bets: A cross-community analysis of three online hacking forums. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2018.

[32] Sergio Pastrana, Alice Hutchings, Andrew Caines, and Paula Buttery. Characterizing Eve: Analysing cybercrime actors in a large underground forum. In *Proceedings of Research in Attacks, Intrusions, and Defenses*, 2018.

[33] Sergio Pastrana, Alice Hutchings, Daniel Thomas, and Juan Tapiador. Measuring ewhoring. In *Proceedings of the Internet Measurement Conference*, 2019.

[34] Sergio Pastrana, Daniel R Thomas, Alice Hutchings, and Richard Clayton. CrimeBB: Enabling cybercrime research on underground forums at scale. In *Proceedings of the World Wide Web Conference*, 2018.

[35] Robert C. Perkins, Marie Ouellet, Christian J. Howell, and David Maimon. The illicit ecosystem of hacking: A longitudinal network analysis of website defacement groups. *Social Science Computer Review*, 41(2):390–409, 2023.

[36] Ildiko Pete, Jack Hughes, Yi Ting Chua, and Maria Bada. A social network analysis and comparison of six dark web forums. In *2020 IEEE European symposium on security and privacy workshops (EuroS&PW)*, pages 484–493. IEEE, 2020.

[37] Srikanth Sundaresan, Damon McCoy, Sadia Afroz, and Vern Paxson. Profiling underground merchants based on network behavior. In *IEEE APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–9, 2016.

[38] Gert Jan Van Hardeveld, Craig Webber, and Kieron O'Hara. Discovering credit card fraud methods in online tutorials. In *Proceedings of the International Workshop on Online Safety, Trust and Fraud Prevention*, 2016.

[39] Anh V Vu, Jack Hughes, Ildiko Pete, Ben Collier, Yi Ting Chua, Ilia Shumailov, and Alice Hutchings. Turning up the dial: the evolution of a cybercrime market through set-up, stable, and covid-19 eras. In *Proceedings of the ACM Internet Measurement Conference*, 2020.

[40] Wei T Yue, Qiu-Hong Wang, and Kai-Lung Hui. See no evil, hear no evil? Dissecting the impact of online hacker forums. *Mis Quarterly*, 43(1):73, 2019.

[41] Xiong Zhang and Chenwei Li. Survival analysis on hacker forums. In *SIGBPS Workshop on Business Processes and Service*, pages 106–110, 2013.

A. Appendix

A.1. Topic Model: First Post

Table 2 shows the results of the LDA topic model for first posts.

A.2. Topic Model: Last Post

Table 3 shows the results of the LDA topic model for last posts.

TABLE 2: Topic model result for first posts

Number	Top 5 Topic Words (with Weights)	Topic Name
0	0.045*"website" + 0.027*"page" + 0.025*"site" + 0.024*"script" + 0.017*"id"	Web
1	0.073*"thanks" + 0.065*"please" + 0.055*"pm" + 0.055*"would" + 0.044*"interested"	General
2	0.028*"file" + 0.017*"app" + 0.013*"files" + 0.011*"key" + 0.010*"use"	Files
3	0.031*"phone" + 0.025*"android" + 0.025*"hf" + 0.021*"video" + 0.019*"youtube"	Mobile
4	0.023*"discord" + 0.020*"account" + 0.018*"buy" + 0.017*"contact" + 0.016*"accounts"	Account (Online)
5	0.057*"good" + 0.042*"pack" + 0.026*"work" + 0.025*"nice" + 0.021*"python"	Pack
6	0.056*"windows" + 0.043*"system" + 0.033*"software" + 0.029*"program" + 0.016*"apps	Software
7	0.091*"code" + 0.013*"number" + 0.011*"return" + 0.011*"string" + 0.010*"bin	Coding
8	0.015*"i'm" + 0.012*"new" + 0.012*"like" + 0.011*"know" + 0.010*"get	General
9	0.060*"help" + 0.058*"need" + 0.049*"please" + 0.040*"account" + 0.038*"link"	Request
10	0.022*"server" + 0.020*"rat" + 0.019*"ip" + 0.019*"use" + 0.014*"help	RAT

TABLE 3: Topic model result for last posts

Number	Top 5 Topic Words (with Weights)	Topic Name
0	0.031*"learn" + 0.027*"looking" + 0.022*"hello" + 0.020*"add" + 0.020*"hacking"	Help
1	0.062*"code" + 0.015*"id" + 0.013*"text" + 0.012*"source" + 0.012*"pi"	Code
2	0.021*"help" + 0.017*"need" + 0.015*"know" + 0.013*"get" + 0.010*"someone"	Request
3	0.124*"discord" + 0.058*"service" + 0.050*"telegram" + 0.037*"software" + 0.035*"windows"	Account (Online)
4	0.018*"good" + 0.016*"like" + 0.012*"time" + 0.011*"people" + 0.010*"one"	General
5	0.050*"account" + 0.027*"accounts" + 0.025*"phone" + 0.023*"need" + 0.021*"hack"	Account (General)
6	0.082*"please" + 0.071*"thanks" + 0.064*"pm" + 0.055*"send" + 0.037*"would"	Request
7	0.036*"bro" + 0.030*"contract" + 0.029*"selling" + 0.020*"bot" + 0.017*"bitcoin"	Market
8	0.043*"file" + 0.018*"open" + 0.018*"version" + 0.017*"download" + 0.010*"pc"	Files
9	0.030*"free" + 0.021*"money" + 0.017*"game" + 0.013*"payment" + 0.011*"video"	Money