Measuring eWhoring

Sergio Pastrana
Computer Science Department,
Universidad Carlos III de Madrid
Leganes, Spain
spastran@inf.uc3m.es

Alice Hutchings
Cambridge Cybercrime Centre
University of Cambridge
Cambridge, UK
Alice.Hutchings@cl.cam.ac.uk

Daniel Thomas
Computer & Information Sciences
University of Strathclyde
Glasgow, UK
Daniel.Thomas@cl.cam.ac.uk

Juan Tapiador
Computer Science Department,
Universidad Carlos III de Madrid
Leganes, Spain
jestevez@inf.uc3m.es

ABSTRACT

eWhoring is the term used by offenders to refer to a type of online fraud in which cybersexual encounters are simulated for financial gain. Perpetrators use social engineering techniques to impersonate young women in online communities, e.g., chat or social networking sites. They engage potential customers in conversation with the aim of selling misleading sexual material – mostly photographs and interactive video shows – illicitly compiled from third-party sites. eWhoring is a popular topic in underground communities, with forums acting as a gateway into offending. Users not only share knowledge and tutorials, but also trade in goods and services, such as packs of images and videos. In this paper, we present a processing pipeline to quantitatively analyse various aspects of eWhoring. Our pipeline integrates multiple tools to crawl, annotate, and classify material in a semi-automatic way. It builds in precautions to safeguard against significant ethical issues, such as avoiding the researchers' exposure to pornographic material, and legal concerns, which were justified as some of the images were classified as child exploitation material. We use it to perform a longitudinal measurement of eWhoring activities in 10 specialised underground forums from 2008 to 2019. Our study focuses on three of the main eWhoring components: (i) the acquisition and provenance of images; (ii) the financial profits and monetisation techniques; and (iii) a social network analysis of the offenders, including their relationships, interests, and pathways before and after engaging in this fraudulent activity. We provide recommendations, including potential intervention approaches.

CCS CONCEPTS

• **Security and privacy** → *Social engineering attacks*; *Social network security and privacy*; • **Social and professional topics** → *Financial crime.*

IMC '19, October 21–23, 2019, Amsterdam, Netherlands

ACM ISBN 978-1-4503-6948-0/19/10…$15.00
https://doi.org/10.1145/3355369.3355597

KEYWORDS

eWhoring, cyber-sex, cybercrime, fraud, underground forums

ACM Reference Format:
Sergio Pastrana, Alice Hutchings, Daniel Thomas, and Juan Tapiador. 2019. Measuring eWhoring. In *Internet Measurement Conference (IMC '19), October 21–23, 2019, Amsterdam, Netherlands.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3355369.3355597

1 INTRODUCTION

Underground forums allow communities to trade illicit material and share knowledge [9, 14, 15, 23, 39, 43]. These forums enable a plethora of cybercrimes, allowing members to easily engage into criminal activities. These include trading virtual items obtained by illicit means, launching DDoS attacks using booter services, or obtaining and using malware [26, 30]. The products and services made available by forum members lower the barrier to entry, enabling those without highly technical skills to engage in deviant or criminal activities [25].

eWhoring techniques have been actively developed since at least 2008. However, until recently the topic remained largely hidden from academic attention. Earlier work [19] provides a qualitative understanding of eWhoring, exploring how it works, how the actors monetise their activities, and possible intervention approaches. This qualitative approach allows us to understand what eWhoring involves and informs the types of measurements we present here.

As with dating scams [7, 18, 29, 38], eWhoring involves social engineering techniques. However, rather than simulating romantic relationships, offenders imitate partners in cybersexual encounters. Targets are asked for money in exchange for pictures, cam shows or even sexual conversations (also known as *sexting*). Initial engagement occurs through chat applications or dating sites. Packs of multiple images and videos of the same person are traded within underground forums. This material is the bait to lure customers into paying for encounters. Underground forums are also used to interchange knowledge and learn new techniques for increasing the benefits obtained from this illicit business.

Contributions. In this work, we provide a quantitative analysis of eWhoring. We start by providing an overview of the steps involved in eWhoring (§2). In §3 we describe the data used for our research, namely information gathered from the underground forums and markets that serve as the basis for initiating, doing, and sharing

knowledge and material for eWhoring. We begin our measurements (§4) by analysing the acquisition and provenance of images. In §5 we estimate the income generated by eWhoring by analysing a subset of forum posts where actors upload evidence of their earnings. In §6, we investigate the roles and other interests of actors involved in eWhoring. Related work is set out in §7. To conclude, we present recommendations and potential disruption approaches and discuss the limitations of our work in §8. The significant ethical and legal risks that were considered when designing and undertaking this research are detailed in the Appendix.

The main findings and contributions of this study are:

- We present a measurement pipeline for downloading, annotating and classifying eWhoring-related material in a semi-automatic way. As this study raises ethical and legal concerns, our pipeline is designed to minimise exposure of researchers to indecent images and enable early detection and reporting of child abuse material. Thus, the pipeline can be applied to research that involves downloading pornographic or potentially illegal images.
- Using a dataset spanning more than 10 years, we perform a longitudinal analysis focusing on three main aspects of eWhoring:
  (1) **Provenance of images**. We found 4k threads in underground forums providing packs of images. From those shared openly and for free, we downloaded around 115k images from cloud storage services. We immediately reported and removed from our servers the 36 images classified as child abuse material. Reverse image searches revealed most were obtained from adult and pornographic sites. We also found images stolen from social networking sites, blogs, photo sharing sites, and online forums, among others.
  (2) **Profits and monetisation techniques**. We analysed 1 868 images posted by 661 actors, allegedly showing their eWhoring earnings, accounting for a total of US$551k. The average reported income is US$774, with some actors reporting more than US$20k. A typical trade of images costs between US$5-50, whereas cam shows are sold for around US$200. PayPal and Amazon Gift Cards are the most used payment platforms for monetising eWhoring. Some actors also use underground forums for money laundering in special boards aimed at currency exchange (e.g., selling Amazon Gift Cards for BTC).
  (3) **Analysis of eWhoring actors**. We analysed 73k actors discussing eWhoring in underground forums. Most of these (~80%) made less than 10 posts whereas 2k actors made more than 50 posts. We use social network analysis to categorise the most popular and influencing actors, and use forum activity to analyse other interests in addition to eWhoring in a set of key actors. We find many actors are initially attracted to the forum's gaming and hacking boards, before engaging in eWhoring.
- Based on our measurements, we provide a set of recommendations and potential intervention approaches to disrupt eWhoring.
- Finally, to make our work reproducible and to foster research in this area, we release our code and part of the processed data publicly: https://github.com/spastrana/ewhoring-analysis. The forum dataset is available from the Cambridge Cybercrime Centre https://www.cambridgecybercrime.uk.

2 BACKGROUND

Previous research into eWhoring used a crime scripting approach, analysing the tutorials and discussions posted on the underground forum *hackforums.net*. These tutorials provide instructions about how to get involved in eWhoring, and how to monetise this activity [19]. The crime script describes the steps involved in eWhoring, from preparation to exchanging the illicitly obtained funds. In this paper we will use the same terminology adopted in the previous research. While not the terms used on the forums, 'actor' refers to those engaging in eWhoring, 'customer' refers to those purchasing, or potentially purchasing images, and 'model' refers to those depicted in the images, with or without their consent.

Typically, images of models are stolen and shared online. Actors provide advice about sourcing images from existing websites, including pornography, social media, and 'revenge porn' sites. Actors prefer a variety of images of the model, including clothed, nude, and 'verification templates', which can be modified (e.g., to display the customer's name). Explicit video footage can be spliced to create customisable interactive cam shows, referred to as 'Video Cam Whores'.

Actors create a unique backstory for each model, such as why they are selling images, and open online accounts. These include email accounts, as well as accounts for communicating (e.g., Skype), accepting payment (usually PayPal or Amazon Gift Cards), and for meeting and attracting customers. The types of websites and applications used to meet customers include chatrooms, video chat, social networks, amateur pornography, classifieds, dating, and online gaming.

Depending on the site, actors will either openly post they are selling photographs and cam shows, or will first start communicating with potential customers before making an offer. At this stage, there may be a process of negotiation and social engineering before a price is agreed on, payment is received and the images are sent. The final step in the process is retrieving the funds from the account they have been deposited. In some cases this involves converting currencies, such as exchanging Amazon Gift Cards for cryptocurrencies.

There are alternative ways of generating income that deviate from the standard eWhoring crime script. These include blackmailing customers, affiliate marketing (e.g., providing images using URL shorteners that first display advertisements), or infecting customers with malware. Customers can also be scammed, with actors not supplying the images they have paid for. Another scam type is 'double dipping', where actors claim there was a problem with receiving payment, and trick the customer into sending the money again.

Our previous work used content analysis methods to generate the eWhoring crime script [19]. This qualitative approach provided useful information about the processes involved in eWhoring, and how this fraud is monetised. While this approach was not quantitative, it provides insights into what would be useful to measure, and

the types of data available. In the following sections we quantify where the images are being obtained from, the characteristics of the actors (such as their other forum activity), and estimate their earnings.

3 DATASET

This study relies on information gathered from underground forums. As in other web forums, users initiate conversations or *threads* by writing an initial *post* and a *heading* which summarises the topic of conversation. Other forum users can post in response to this thread. The forums are made up of numerous *boards*, which contain threads relating to specific topics.

We use the CrimeBB dataset [27] which contains data scraped from 15 different underground forums, and is available from the Cambridge Cybercrime Centre[1]. One of the forums is *Hackforums*, the largest English-language underground forum. *Hackforums* has been operating since 2005 and contains a board specifically for eWhoring. Other forums, such as *blackhatworld.com* allegedly prohibit conversations about eWhoring, since discussions about 'unlawful activities, fraud, or deception' [5] contravene their terms of service. However, as shown in Table 1, forum administrators have failed to remove such conversations.

To gather conversations related to eWhoring, we searched for two specific keywords (i.e., 'ewhor' and 'e-whor') in the headings of all the threads contained in CrimeBB, (comparison was done in lowercase). Since the role of headings is to summarise what the thread is about, eWhoring related threads will in most cases contain either one of these two keywords. While some threads might be missed, this in unlikely and our analysis should not be affected. We also include all the threads from the specific board dedicated to eWhoring in *Hackforums* (more than 36k threads at the time of writing). Overall, our analysis leverages 44k threads and 626k posts made in 10 underground forums by more than 72k actors. The data spans more than 10 years of activities: the first post in the dataset was made on November 2008 and the last on March 2019.

Table 1 shows the number of threads, posts, and actors that relate to eWhoring for each forum. The largest eWhoring community is found on *Hackforums* (which was expected since it contains a dedicated section for eWhoring and is frequently referred from other forums). *OGUsers* is a community focused on trading online accounts with popular or interesting names. Here, we observe most of the threads related to eWhoring are for trading chat-related accounts (e.g., Snapchat or Kik) with feminine names.

Limitations. Using data from underground forums provides a single picture of the landscape. These forums are platforms where users initiate in deviant activities, such as eWhoring. They are also used for sharing knowledge, tools and material required for such activities. Thus, they are an interesting source of information, and allow for analysing the material used for eWhoring (such as images and packs), the earnings reported in the forums, and also the social network of forum actors. Also, we only rely on data publicly available from forum conversations. For example, we neither use information from private messages to build the social network, nor analyse packs sold in the forums. Moreover, we do not analyse data from platforms where eWhoring is actually carried out, such as adult chats or social networks. Finally, we have not access to data from endpoint payment platforms, such as Amazon or PayPal, and thus our estimation of profits is based on data publicly reported by actors. While this data could be deceptive, we believe this is unlikely since, apart from bragging rights, there are few other incentives to falsify the data. Thus, the results presented in this paper must be seen as an initial approximation to this previously-unexplored activity.

[1] https://www.cambridgecybercrime.uk/

Forum	#Threads	#Posts	First post	#TOPs	#Actors
Hackforums	42 292	596 827	11/08	4 027	64 035
OGUsers	1 744	23 974	04/17	76	5 586
BHW	258	2 694	04/08	0	1 420
V3rmillion	95	1 348	02/16	6	697
MPGH	62	922	07/12	12	341
RaidForums	48	405	03/15	10	318
Others (4):	21	614	05/15	6	586
TOTAL:	44 520	626 784	11/08-03/19	4 137	72 982

Table 1: Number of eWhoring related conversations per forum. (TOPs = Threads Offering Packs.)

4 IMAGE PROVENANCE

One of the key requirements for successful eWhoring is to use a good set of images and videos from the same model, known as 'packs'. Good packs are those containing 'unsaturated' material, i.e., which have been barely or never used by others and thus are less prone to raise the suspicions of the customer or be blocked from the site being used to attract traffic. While some actors may create their own packs, we observe ready-to-use packs are frequently shared and sold in underground forums (indeed, we suspect the tutorials providing information about how to monetise eWhoring are designed to increase demand for these packs). These packs contain images from the same (or visually similar) model at the various steps of a 'fake' encounter, including dressed, nude and sexual images and videos.

We are interested in analysing where these images are obtained from and how they are shared between actors. We developed a pipeline (see Figure 1) to analyse the images that takes into account ethical and legal concerns (see Appendix). Using our pipeline, we: (1) retrieve the packs shared in the forums; (2) download the images; (3) filter out images related with child abuse; (4) automatically classify images containing explicit, sexual, or nude content as 'Not-Safe-For-Viewing' (NSFV), to avoid viewing them manually; and (5) use a reverse image search to identify the domains where these images have been obtained. In the following subsections we describe the various steps involved in our pipeline, including the limitations and results.

4.1 Extracting Threads Offering Packs

The first step in our pipeline is to identify those threads where packs are offered, dubbed 'Threads Offering Packs' (TOPs). During preliminary inspection of the dataset, we noted most TOPs provide previews, i.e., one or more samples of the images contained in

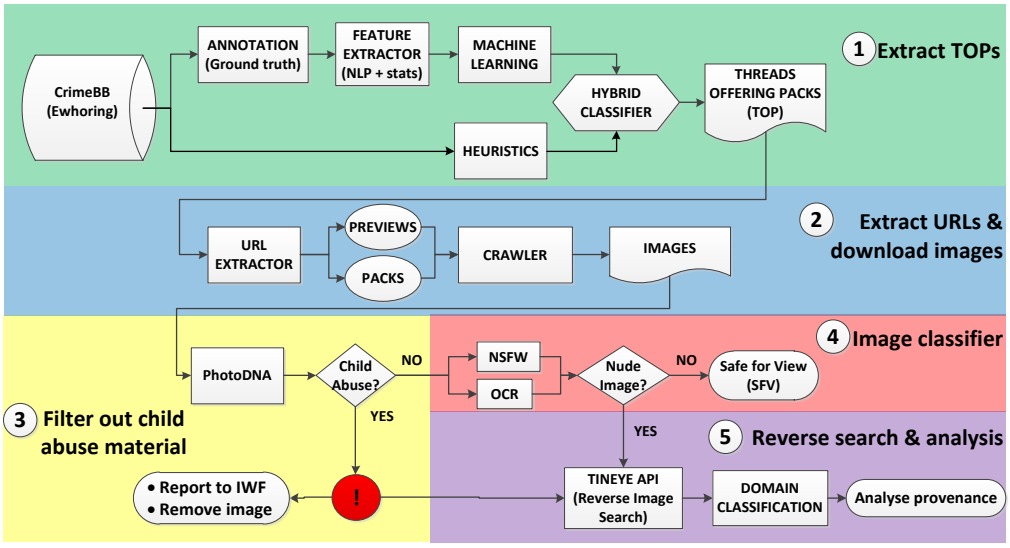

Figure 1: Pipeline used to identify, download and analyse packs of images used for eWhoring.

the pack. These previews are shared using online image sharing communities such as *imgur* or *Gyazo*. The actual packs are shared using external cloud storage platforms such as *MediaFire* or *mega.nz*. TOPs are typically popular threads with several replies, mostly in cases where these are shared only upon explicit request.

We use a hybrid classifier, which relies on a machine learning (ML) algorithm and a set of heuristics extracted from the thread headings and posts. The ML algorithm is fed with a set of statistical features. Concretely, for each thread it extracts: the number of replies; the number of links to cloud storage and image sharing sites, and number of links to other threads in the forum; the length of the first post; and a set of features extracted from the text using natural language processing (NLP). For the latter, we parse thread headings and posts into a document-term matrix to get word-counts. We strip punctuation, convert to lower case characters, ignore numbers and exclude stop words. Finally, these word counts are transformed using TF-IDF ('term frequency inverse document frequency'). Additionally, the feature set used for the ML algorithm includes the number of special keywords and characters in the thread headings, such as question marks, keywords related to selling/buying (e.g., 'WTS', or 'looking for') and keywords related to tutorials and mentoring (e.g., 'Guide' or 'Tut'). Including these features helps the ML algorithm to discard threads *asking for* instead of *offering* packs. See Table 2 for the complete list of keywords used in our methodology.

The heuristics are formed through our expertise in analysing forum data [19, 25, 27]. Concretely, for each thread we account for keywords frequently observed in TOP headings such as 'images', 'video' or 'unsaturated'. As with the ML algorithm, we also account for both the number of question marks and the presence of keywords related to buying to discard threads asking for packs. If either method classifies a thread as offering packs, this is included in our pipeline to extract links.

A single annotator labelled a subset of 1 000 threads from the various forums. Then, 800 threads were used to train a classifier and 200 threads for testing. We use the Linear-SVM algorithm since it offered the best results in previous experimentation with our dataset [8]. We evaluate the classifier using standard metrics for information retrieval, i.e., precision, recall, and F1 score.

Results. From the set of 1k annotated threads, 175 correspond with TOPs. Using this dataset, the hybrid classifier achieves a precision of 92%, recall of 93% and F1 score of 92%. While the results could be improved, the classifier is helpful to extract a small subset from a large dataset of threads.

In total, we are able to extract 4 137 TOPs. The ML classifier extracted 3 456 TOPs and the heuristic-based classifier 2 676 TOPs. Out of these, 1 995 were extracted by both approaches. This show the benefits of using an hybrid approach: on the one hand, the heuristics can help to automate the search of TOPs with known characteristics. On the other side, the ML classifier can learn new patterns from TOPs and apply those to the unclassified dataset. The third column in Table 1 shows the number of TOPs found per forum. Interestingly, we observe there are no TOPs on *BlackHatWorld*. By visually inspecting the headings of the 258 threads we observe most are discussing various aspects of eWhoring, including that it is banned in the forum. There are also some tutorials or ebooks provided and requests for pictures. We conclude that, while the rules ban any eWhoring-related conversation, threads providing images (i.e., packs and previews) are being removed.

Limitations. The feature set used for the ML algorithm includes variables extracted from text using NLP techniques. Previous work have shown the limitations of applying such techniques in underground forums data, e.g., due to the use of specific jargon, misleading vocabulary or syntax and grammar errors [8, 12]. A potential solution would be to normalise the data into a common format. In our pipeline this limitation is partially addressed since the ML algorithm combines NLP variables with other statistical features. Moreover, we enrich our classification using heuristics derived

Extract eWhoring-related threads	*ewhor, e-whor*
Classify Thread Offering Packs (TOPs)	*pack, packs, package, packages, pics, pictures, videos, vids, video, collection, collections, set, sets, repository, repositories, selling, wts, offering, free, unsaturated, new, giving, compilation, private, girl, girls, sexy.*
Detect info-requesting posts	*[question], [help], need advice, need, needed, wtb, want to buy, req, request, question, looking for, give me advice, quick question, question for, i wonder whether, i wonder if, im asking for, general query, general question, i have a question, i have a doubt, help requested, how to, help please, help with, need help, need a, need some help, help needed, i want help, help me, seeking*
Detect threads providing tutorials	*tutorial, [tut], howto, how-to, definite guide, guide*
Extract posts sharing earnings	*earn,profit,money,gain*

Table 2: Keywords used during our methodology

from our experience in analysing forum data related to eWhoring [19]. We observed most TOPs include specialised keywords such as 'unsaturated' or 'pack'

4.2 URL extraction and crawler

Using regular expressions we extract URLs from the content of each extracted TOP. Using a whitelist of known image sharing sites (used to share previews of the packs) and cloud storage services (hosting the packs) we extract the links corresponding with both previews and packs. This whitelist is compiled using a snowball sampling technique: starting with a known set of domains, we parse all URLs extracted from the TOPs, and manually analyse a subset of the domains that do not belong to the whitelist, visiting their landing sites. This process is repeated until the URLs are either unknown or do not belong to cloud storage services or image sharing sites. Finally, the list of links corresponding to previews and packs is fed to a custom crawler which downloads the images and (if needed) decompresses the packs into folders.[2] For each link, we also annotate associated metadata (e.g., the post identifier and author).

Results. Tables 3 and 4 show the number of links extracted per image sharing site and cloud storage service respectively. *Imgur* is by far the most popular platform to share previews, followed by *Gyazo* and *ImageShack*. An inspection into their terms of service reveal these platforms forbid uploading images containing nudity or violating copyright. Of the cloud storage services, *MediaFire* is by far the the most used platform, followed by *mega*, *Dropbox*, and *oron* (a now defunct site). Again, the terms of service forbid content that 'violates someone else's individual rights or copyright' (*MediaFire*). Likewise, *mega*'s terms specify 'You are strictly prohibited from using our services to infringe copyright' (mega). It appears *oron* closed down after being sued for copyright infringement (including pornographic material) [20].

We crawled the links and were able to download 5 788 images from image sharing sites and 111 288 images contained in 1 255 packs. Many files and images had been deleted. After removing duplicates (for example, 127 images were found in at least 20 different packs), there were 53 948 unique files. As these packs are offered at no charge, and thus are likely 'saturated' (i.e., they have been used for eWhoring before) we had expected to observe duplicate images. We also found that not all files downloaded from image sharing sites correspond with actual previews, which we will discuss in §4.4.

Site	#Links
imgur	3 297
Gyazo	1 006
ImageShack	679
prnt	383
photobucket	311
imagetwist	105
imagezilla	97
minus	51
postimage	47
imagebam	44
Others	700
Total	7 314

Table 3: Number of links per image sharing site.

Site	#Links
MediaFire	892
mega	284
Dropbox	130
oron	95
depositfiles	46
filefactory	37
drive.google	31
ge.tt	28
zippyshare	25
filedropper	24
Others	94
Total	1 719

Table 4: Number of links per cloud storage service.

Limitations. Some of the packs are released to users who reply to a given thread or pay a fee. Due to ethical concerns we do neither of these, and thus we are not able to download such packs. Thus, our results are limited to packs openly shared for free. Concretely, out of the 4 137 TOPs, we were able to extract links from 774 threads (18.71%). Also, we look for links leveraging a whitelist of cloud storage and image sharing sites. Accordingly, we might be missing some sites. To reduce this limitation, we used a snowball sampling method. When downloading images, we encountered two limitations. First, we did not download packs from some sites requiring registration, e.g., Dropbox or Google Drive, where crawling violates their Terms of Service. Second, many of the cloud storage services and image sharing sites are either defunct or restrict the lifetime of the links for free or trial accounts. Thus, while we have retrieved links posted in the past, we were unable to download the content.

4.3 Filtering out child abuse material

A potential legal issue for actors involved in eWhoring is downloading and distributing child abuse material. As we are downloading images to our servers, this is also a concern for our research. Accordingly, after due discussion with our Research Ethics Board (see Appendix), we contacted the UK Internet Watch Foundation (IWF)[3] to explain our study and ask for their assistance. As part of their

[2] Interested readers can get details on the crawler by inspecting the source code at our repository: https://github.com/spastrana/ewhoring-analysis/tree/master/tools/crawler

[3] https://www.iwf.org.uk the UK's INHOPE hotline operator.

cooperation, we were granted access to the PhotoDNA Cloud Service [22], which computes a hash of a given image and matches it against a database of known child abuse material. These images have been tagged due to clearly containing child sexual content or have been reported as depicting someone who is underage. Each image matching the PhotoDNA list was immediately reported to the IWF and deleted from our servers. We also reported the URLs of other sites where these images were located, obtained from the reverse image search (see §4.5).

Results. We found 36 images matched the PhotoDNA hashlist. Among them, the IWF actioned 61 URLS, 60 related to a single UK victim aged 17, and one related to a 7 to 10 year old victim. While the rest of the images match the hashlist due to other organisations grading them as child abuse material, these were not actionable by the IWF since they were not able to verify the age of the persons depicted. Regarding the 61 URLs actioned by the IWF:

- **Severity**. 20 of the images were category A (images involving penetrative sexual activity; images involving sexual activity with an animal or sadism), 36 Category B (images involving non-penetrative sexual activity), 5 Category C (other indecent images not falling within categories A or B).
- **Hosting location**. One site was hosted in the UK (taken down by the IWF), 30 were in North America (USA and Canada) and 30 were hosted in other European countries.
- **Site types**. All recorded as being on 'Free Hosting' sites over 36 separate domains, namely: 26 image sharing sites, 9 forums, 3 blogs, 2 social networks, 1 video channel, and 20 regular websites.

As possession of child abuse material is a crime in many jurisdictions, actors and customers downloading these packs are placing themselves at risk of criminal charges. In some jurisdictions, these are 'strict liability' offences, therefore not intending to access images of children would not necessarily be a legal defence. A defence may be available if the defendant can prove innocence, such as having had immediately reported the images to the police and deleted them (as we did), although in other cases they may have to rely on prosecutorial discretion not to pursue charges [10].

The images were downloaded from links posted in 36 different threads which were replied to by 476 different actors. An inspection of the replies show most were expressed gratitude and recognition such as *'Downloading, thanks for the share!'* or *'just download the pack, amazing pack'*. This indicates at least 476 actors are potentially downloading child abuse material. This is a lower bound as many users may have downloaded the packs without posting a response. We also observed some cases where users discussed the age of the models in the packs, which suggests the community is aware of the potential legal risks associated with downloading child abuse images. For example, one reply exclaimed:

> *'you have to take the image down. She is 100% under age, just look at her!! And thanks for the share anyway'*

Limitations. The detection of images containing child abuse material is limited by the accuracy of the PhotoDNA technology from Microsoft. While offenders might try to evade detection by modifying the images, PhotoDNA leverages Robust Hashing to detect images that have been modified, e.g., using compression algorithms or geometric distortions [37]. To the best of our knowledge, the robustness of this tool has not been independently verified, and doing so is forbidden by the terms of use and out of the scope of our study. However, this is the state-of-the-art technology used by law enforcement and non-profit organisations in the fight against online child exploitation, including the IWF and the US National Center for Missing and Exploited Children (NCMEC).

4.4 Image classification

To address some of the ethical and legal concerns raised by our study (such as inadvertently viewing child abuse material, and the potential for psychological harm – see Appendix), we designed an approach to minimise the amount of indecent images visualised by the researchers. Accordingly, we developed a NSFV classifier which indicates whether an image contains indecent content or not. Recall that our classifier is used to discern between indecent images from models, possibly showing explicit sexual content, and other images containing text, such as screenshots of payment platforms or chats. Therefore, in addition to the images downloaded from TOPs, we also apply the classifier before the manual inspection of images related with eWhoring earnings (see §5).

Before identifying the provenance of images, we filter out those not actually depicting models. Images contained in packs typically contain a set of pictures from the same (or visually similar) model in various stages of a virtual sex encounter, i.e., dressed (normally in a suggestive manner, which are used for attracting customers), naked (partially or full), or involving sexual activity. This also applies for pack previews. However, our link extraction for image sharing sites may retrieve links to other types of images, such as 'proof-of-earnings' (images showing the eWhoring earnings reported by forum actors, typically screenshots of dashboards from payment platforms) or other screenshots (e.g., of a chat discussion between an actor and a customer). As these types of images usually contain characters, the NSFV classifier combines the output of a detector of nudity in images with an Optical Character Recognition (OCR) classifier.

For the nudity detector we used Yahoo's OpenNSFW classifier [21] (where NSFW stands for 'Not-Safe-For-Work'), which leverages a model trained using deep learning and provides a probability score of an image containing indecent content. For the OCR classifier, we use the Tesseract software [32], which outputs the number of words recognised in an image. Using these two scores (NSFW and OCR), we developed a set of heuristics to determine whether an image is NSFV or not. These heuristics combine thresholds from both classifiers in a set of rules. The heuristics are tuned using both a validation dataset of 180 labelled images (including sexual and non-sexual content) released by Lopes et al. [2] and a set of 60 images manually retrieved from the web with textual content (e.g., documents, bills, source code, etc.) and without textual content (including landscapes, screenshots of virtual games, or pictures taken from random people) .

Results. We have empirically tested various combinations of the scores provided by the Yahoo NSFW and OCR classifiers against the validation dataset. In general non-nude images receive a NSFW score lower than 30%. There is an exception with images of clothed models with high proportion of human body, since these usually

have a NSFW score which is between 10% and 70%. In our implementation, the main task is to discern between images from models (packs and previews) and images showing textual content (e.g. showing earnings, chats, etc.). Thus, we rely on the OCR to fine-tune the scores of the classifier. Concretely, images including text and characters (e.g. source code, memes, or screenshots) are properly recognised by the OCR classifier, which extracts the text included in the image. After empirically testing various thresholds by running the classifier on the validation dataset, we established the heuristics shown in Algorithm 1, which best classify the validation dataset. These heuristics are conservative, since they resulted in a 100% detection of NSFV images (meaning that all images tagged as nude are detected as NSFV) while having few false positives (nearly 8%).

Algorithm 1: Classify images into SFV or NSFV

```
Input: Image
Output: True if SFV, False otherwise
NSFW ← openNSFW(image)
OCR ← tesseract(image)
if NSFW < 0.01 then
    return True
else if NSFW > 0.3 then
    return False
else if NSFW < 0.05 then
    return OCR > 10
else
    return OCR > 20
```

Among the 5 788 images downloaded from image sharing sites, 3 496 were classified as NSFV and thus we include them in the set of 'previews'. Other links either pointed to error messages (e.g., 'This image violates our Terms of Use and has been removed from view') or screenshots showing the directories of the packs, including image thumbnails.

Limitations. We rely on open-source tools to classify images. Yahoo's OpenNSFW provides a trained deep learning-based model which provides a NSFW score for each input image. While the definition of NSFW is subjective, this model suits our needs. Indeed, as stated by the authors, this model 'can be used for the preliminary filtering of pornographic images' [21], which is precisely our goal. Moreover, we enrich the classification by using the OCR classifier, using the Tesseract OCR engine [31, 32]. Again, we rely on third party software whose exact accuracy is not known to us. However, Tesseract is one of the most used technologies for recognising characters in images, both in industry [35] and academia [31].

The threshold and heuristics for the classification are established in a semi-automatic process which is tuned using a dataset with few images (240). With ethical concerns in the forefront, we select those thresholds in a conservative way to minimise false negatives (i.e., including indecent images in the set of SFV images). This increases the likelihood of false positives, which might affect the completeness of our results. In this regard, false positives are images that, not being from models, are classified as so. These images might be included in our final dataset of images for which we conduct reverse image search, possibly biasing our final results. The most problematic (i.e., hard-to-classify) pictures are those that, not containing nudity, are tagged as NSFV due to: i) not having any text, and ii) containing colours or textures resembling the human body. Additionally, a recent paper shows how adversaries might modify images to prevent detection of pornography [40]. Due to the lack of a labelled dataset for eWhoring images, we can not test the actual performance of our classifier, and this might be viewed as a filter for indecent images. However, since our classifier is used to discern between model previews (usually containing nudity and none or few characters) and images having high percentage of text (e.g. screenshots from chats or payment platforms), the number of images wrongly classified is likely negligible. Indeed, during manual annotation of 2 067 images showing earnings (see §5) we have not visualised any image from models.

4.5 Reverse image search and domain classification

To analyse the actual provenance of the images being used for eWhoring, we use the image reverse search service provided by TinEye [36]. This service compares a given image against a database of more than 29 billion images crawled from the web. Each comparison outputs a similarity score, and if this score is greater than zero it is considered a match. If one or more matches are found, a report is created indicating for each match, among other information: i) the domain and URL where the image is (or was) hosted; ii) the backlink from where it was crawled and; iii) the crawling date. Additionally, to analyse whether the images were online before they were posted in the forums, we have used the Wayback Machine [3] to explore the Internet Archive [4] for each of the matching URLs.

We performed reverse image searches for the entire set of 3 496 preview images classified as NSFV. Due to the large number of images included in the packs (around 111k), we selected 3 images per pack for reverse searching. We select the images with the lowest, median and highest NSFV score from each pack. As these correspond to the same model, we assume they have been taken from the same site. In total, we performed reverse searches for 3 644 images from the 1 255 packs (note some packs have less than 3 images).

We used domain classification tools to gain additional knowledge of the type of sites offenders rely on to collect material for their packs. We use three well-known classifiers: Cisco's OpenDNS[5] domain tagging service, McAfee's URL ticketing system,[6] and VirusTotal's URL reputation service.[7]

Results. Table 5 shows the results of the reverse image search performed using TinEye's service. In total, we got 3 621 responses for the images in the packs plus all (3 496) the preview images. From these, we got matches for 74% and 49%, respectively. This difference might be due to modifications made on previews, e.g., by adding a watermark or shadowing parts of the images. Actors purposely modify these images to bypass reverse image searches [19]. However, the ratio of matches per image is substantially higher in the case of previews (observed on average in more than 17.3 sites), as opposed to packs (which are shown on average in 12.7

[4] http://archive.org
[5] https://www.opendns.com
[6] https://www.trustedsource.org
[7] https://www.virustotal.com/gui/home/url

	Total	Matches	Seen Before	Ratio	Max
packs	3 621	2 675 (74%)	2 011 (55.54 %)	12.7	642
previews	3 435	1 683 (49%)	1 340 (39.01 %)	17.3	1 969

Table 5: Number of matches for the reverse image search. Seen Before means that the image was URL was online before the image was posted in the forum. Max and ratio refers to the maximum and average number of matches per image.

sites). The column 'Seen Before' in Table 5 shows the number of matches whose URL has been crawled (either by TinEye or the Wayback machine) before the corresponding image was posted in the forum. Thus, these images were online before they were shared as packs. For the remaining matches (i.e. those whose crawling date is posterior to the post in the forum) we cannot claim whether they were online before or not.

From the set of 1 255 packs analysed, 203 are *zero-match*, i.e., composed by images with zero matches in the reverse image search. We observed various actors offering several zero-match packs, with a single actor sharing 47 zero-match packs (out of 100 shared packs). Zero-matches might be due to packs composed by pictures: i) of actual models, ii) obtained from sites which are not included in the TinEye database, or iii) modified to bypass reverse searches (e.g., by mirroring the images). The last option can be easily performed using automated tools, which are shared in underground forums [19].

The reverse image search resulted in 5 917 different domains. The distribution of categories provided by each domain classifier has a long tail, with around 4-5 categories accounting for more than 50% of the domains (see Table 6). The top categories are mostly porn-related sites, though the distribution is quite different depending on the classifier. Our results suggest images are also taken from a variety of sources other than porn/nudity sites, including social networks, online shops, photo sharing services, blogs and online forums/discussion boards, among others. This is not surprising given how rich in multimedia material such sites are, especially for clothed images used during the early stages of engagement with customers. This range of users whose personal pictures might be illicitly used in eWhoring activities could be wider than initially expected, since it seems to encompass more than models in the porn industry.

Limitations. We rely on TinEye for the reverse image searches. TinEye's matching engine claims 'to deal with a broad range of image transformations, including resizing, cropping, edits, occlusions and colour changes, amongst others' [36]. Each image is tested against a database, which at the time of writing contains 29 billion images. Nonetheless, our results are biased by the effectiveness in dealing with image modifications, e.g., mirroring or shadowing, and also by the completeness of the database (e.g., images obtained from private social network profiles might not be indexed). Thus, our results are also limited by the scope of the reverse search. Finally, we note that both TineEye and the Wayback machine might be incomplete, so an image recorded (crawled) after it was posted in the forum does not imply that it was not available before. It would be needed to get actual timestamps from the pages stating otherwise, but that seems unlikely with automated processing (and due to ethical and technical reasons we did not manually visit the URLs.)

As for the domain classification tools, we follow the same approach used by previous work (e.g., [28]) to perform URL categorisation. Despite this, our results might be biased since the accuracy of such services is generally unknown. Other limitations include: (i) the vagueness of some of the categories, which makes it impossible to conduct a thorough analysis, (ii) the lack of classification for some domains, which is quite large in the case of OpenDNS (22%), and (iii) the potential discrepancies between the categories provided by different services.

5 FINANCIAL PROFITS

Underground forums bring together actors interested in easy money making methods.During our analysis of underground forum data, we observed eWhoring actors boasting about their earnings for a variety of purposes. In most cases, it appears this is aimed at attracting new actors or for comparing incomes, with threads titles such as '*Post your earnings*' or '*How much you make?*'. Some users regularly post in response to these threads, providing their earnings on a daily or weekly basis. Others share their overall earnings.

In other cases, earnings are shared to promote a service (e.g., mentoring or teaching) or products (e.g., e-books or packs of images). While some users post their incomes without providing proof, this is considered untrustworthy and is not accepted by the community. Thus, most users provide proof-of-earnings in the form of images showing payments received, such as screenshots showing the dashboard of the platform used (typically PayPal and Amazon Gift Cards) that contain account balances and transactions; confirmation emails; receipts; or even pictures of cash. Similar to the previews used to promote packs of images (see §4), proof-of-earnings are uploaded to image sharing sites such as *imgur* or *Gyazo*.

We analysed 2k of these proof-of-earnings images shared in underground forums. We first present our methodology to retrieve and process the images. We then provide the results of our analysis, including the types of payment platforms used and the reported earnings over time and by actor. We recall that these images can be falsified, e.g. to inflate earnings and brag about their capabilities, and thus our results must be seen as an approximation to actual earnings.

5.1 Measurement pipeline

To estimate the income generated by eWhoring, we developed a methodology for downloading and extracting information from proof-of-earning images. First, we extracted posts related with the sharing of earnings using a set of heuristics. We searched for eWhoring related threads containing the words 'you make' or 'earn' in their heading. We also included eWhoring threads from a board in *Hackforums* called 'Bragging Rights' (used to discuss earnings from a variety of underground activities). This query yielded 1 084 threads, from which we extracted 725 posts containing links to image sharing sites. We also queried the dataset for posts containing the keyword 'proof' with trading-related terms (see Table 2). This way, we obtained a further 551 posts containing links to image sharing sites. We applied regular expressions to extract the URLs.

McAfee category	#Domains	Distrib. (%)
Pornography	2078	28.75
Blogs/Wiki	781	39.55
Entertainment	466	46.00
no_result	361	51.00
Forum/Bulletin Boards	208	53.87
Online Shopping	207	56.74
General News	187	59.32
Provocative Attire	181	61.83
Marketing/Merchandising	169	64.17
Games	166	66.46
Internet Services	155	68.61
Media Sharing	141	70.56
Dating/Personals	126	72.30
Portal Sites	114	73.88
Parked Domain	112	75.43
Malicious Sites	109	76.94
Social Networking	102	78.35
Business	102	79.76
Nudity	88	80.98
PUPs	87	82.18
Humor/Comics	83	83.33
Streaming Media	71	84.31
Illegal Software	70	85.28

VirusTotal category	#Domains	Distrib. (%)
adult content	2136	17.78
porn	1510	30.35
sex	1465	42.55
uncategorised	1367	53.93
business	530	58.34
no_result	368	61.41
entertainment	283	63.76
shopping	272	66.03
news	247	68.08
news and media	226	69.96
blogs	190	71.55
onlineshop	172	72.98
education	160	74.31
business and economy	157	75.62
message boards and forums	154	76.90
information technology	153	78.17
computers and software	147	79.40
games	138	80.54
social networking	128	81.61
onlinedating	118	82.59
parked	115	83.55
marketing	105	84.42
sports	78	85.07

OpenDNS category	#Domains	Distrib. (%)
Pornography	2054	24.56
no_result	1901	47.29
Nudity	1761	68.34
Adult Themes	405	73.18
Lingerie/Bikini	286	76.60
News/Media	201	79.01
Blogs	154	80.85
Ecommerce/Shopping	119	82.27
Forums/Message boards	109	83.57
Photo Sharing	100	84.77
Sexuality	99	85.95

Table 6: Top categories (85% of the distribution) for the domains obtained through reverse image search. The percentages refer to the total amount of tags, not domains, since a domain classifier can provide more than one tag per domain.

In total, we extracted 2 694 unique URLs from 1 276 posts. Most of these (89%) are from *Hackforums*.

Using our crawler described in §4.2, we downloaded a total of 2 366 images. While our aim was to download proof-of-earnings, the images downloaded could also contain indecent content (e.g. pack previews). Accordingly, we followed the same precautions taken when downloading the packs and previews. First we applied the NSFV classifier to avoid visualising indecent images, and the IWF hashlist to detect child abuse material. After filtering out indecent images (no child abuse material was found), we were able to analyse 2 067 images.

We manually analysed the images and found 199 were not proof-of-earnings. These were, for example, screenshots of chats with customers, error images, or banners. Thus, we found a total of 1 868 images (78.9% of the total downloaded) showing earnings.

We annotated the images with the following information: payment platform (e.g., Paypal, Amazon, etc.); currency (GBP, USD, EUR, etc.); number of transactions; time span; total amount; language; and any other relevant information (such as notes left by the user).Some images included transactions in multiple currencies, which were manually converted to the most frequently used currency included in the proof-of-earning. We then converted automatically all rates to USD. In both cases, we use a historical exchange rate list to get the corresponding rate when the transaction was performed.

Additionally, we analysed the threads started by eWhoring actors in the *Currency Exchange* board on *Hackforums*. This board is used to transform earnings into different virtual currencies, e.g., from PayPal to BTC. Thus, it is an indicator of the financial activity of forum actors. Most of the threads in this board use a *de-facto* standard format where the currency offered follows the tag *[H]* and the currency wanted follows the tag *[W]*. We leverage this format to automatically identify the currencies traded by eWhoring actors. To restrict our analysis to activities related to eWhoring, we only include Currency Exchange threads from actors who have write more than 50 posts in eWhoring-threads.

5.2 Analysis

Overall, there are 661 actors that posted proof-of-earnings, totalling more than US$511k. The average reported per actor is US$774. As shown in the left chart in Figure 2, most actors report earnings less than US$1k. Either they made little money, or they have not shared all their earnings. Indeed, the actors reporting the highest profits are those who share their earnings regularly. The right chart in Figure 2 shows that, in general, actors reporting more earnings provide more proof-of-earnings. For instance, more than 50% of actors reporting more than US$5k have posted 8 or more images. We observe a single actor who posted 46 images, reporting earnings of US$18 097.12 between March 2017 and December 2018. Note that actors sharing earnings account for around a 30% of the actors that have made more than 50 posts, and those providing proofs might show only a small proportion of their actual earnings. Also, some images could be falsified to bloat earnings. Thus, our results must be viewed as an estimation of the actual profits generated by eWhoring.

Of the 1 868 images, around 60% (1 116) show detailed information about the incoming transactions (i.e. dates and amounts). From these, we calculate the average amount per transaction is US$41.90. Some of the proof-of-earnings are accompanied with notes indicating the source of the transactions. From these, we observe most of the transactions are between US$5 and US$50, and usually correspond with the trading of a few set of images. Larger transactions (of US$200 or more) are payments for video encounters or extended periods of time. We also observe earnings shared to promote services or goods, such as mentoring or ebooks.

Amazon Gift Cards and PayPal are by far the most used payment platforms used for eWhoring, with 934 and 795 images respectively.

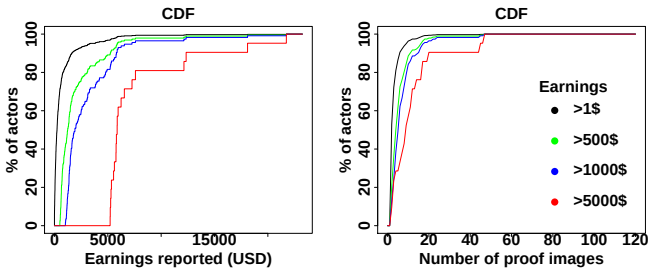

Figure 2: Cumulative frequencies of earnings (left) and number of images (right) posted by actors.

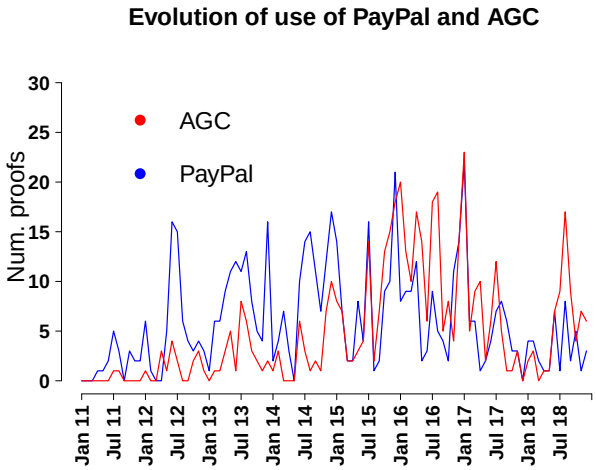

Figure 3: Evolution of number of proof-of-earnings using Amazon Gift Cards (AGC) and PayPal.

We also observe 35 images showing Bitcoin exchanges. The exchange of Amazon Gift Cards on underground forums has increased in recent years [27]. We observe similar patterns in eWhoring. Figure 3 shows the number of images using Amazon Gift Cards and PayPal per month. Since 2016, Amazon has become the preferred payment platform for eWhoring.

We also analysed 9 066 threads in the Currency Exchange board made by 686 eWhoring actors. Recall we only consider in this analysis those actors with more than 50 posts related to eWhoring, and only include threads in Currency Exchange made after the actors started in eWhoring. Table 7 shows the number of threads asking for and offering Amazon Gift Cards (AGC), PayPal and Bitcoin (the top 3 currencies on offer). There is a notable difference between Amazon Gift Cards offered (1 498) and wanted (310). Bitcoin is the most wanted currency, suggesting actors use the Currency Exchange board for exchanging eWhoring profits into Bitcoin. We assume the Bitcoin on offer are not eWhoring profits, as these are relatively uncommon. Indeed, the eWhoring community discourages requesting payment in Bitcoin, as customers are unlikely to hold this currency [19].

Currency	PayPal	BTC	AGC	?	others	Total
Offered	3 707	2 763	1 498	839	259	9 066
Wanted	2 801	4 626	310	1 128	201	9 066

Table 7: Number of threads offering and asking for currencies by forum actors with more than 50 posts in eWhoring. '?' means unclassified.

6 ANALYSIS OF EWHORING ACTORS

In this section we analyse the social relations and features of actors involved in eWhoring. We first present the techniques used for analysing the actors who are involved and then provide a general overview of them. We then focus on the key actors, namely those who are particularly interesting due to their characteristics and the types of activities they are involved in. This part of the study focuses on *Hackforums*, the underground forum containing the largest community for eWhoring (see Table 1). We use previous, current and posterior forum activity to analyse the pathways followed by key actors, including the other interests they exhibit during their interactions on the forum.

6.1 Social network and feature extraction

To analyse the social interactions between forum members, we built a network from the public conversations of members in the forum, i.e. who responded to whom in the threads. We consider actor A has responded to actor B if either A explicitly quotes a post made by B in a reply or if A directly posts a reply in a thread initiated by B, without quoting any other post. We account for all the interactions between forum members in threads related to eWhoring. We built a social graph where nodes correspond with forum actors and edges are the interactions between them, weighted by the number of responses. We also computed popularity metrics based on the replies to threads initiated by actors. These include a H-index (a metric widely use to measure popularity of scholars, which indicates that an actor has H threads with at least H replies), and the i-10, i-50 and i-100 indices (i.e., the number of threads with at least 10, 50, or 100 replies).

Additionally, for each actor we get the date when the first and last eWhoring-related posts were made, as well as the registration date and the date of the last activity in each forum. This allows us to analyse previous and posterior activity of the actors in the forum. We also account for the total number of posts made in both eWhoring and other sections to analyse whether actors are exclusively using the forums for eWhoring, or if they are also interested in other boards. We note some other forum activities might still be related to eWhoring, e.g., actors might post in Currency Exchange boards to cash out their eWhoring earnings.

To analyse the interests of forum users, we follow a similar approach used in previous work [25, 27]. We leverage various categories defined in *Hackforums* (e.g., Hacking, Coding, Marketplace, etc.), and then construct the interest of a user A in a category C by counting the number of posts and threads made by A in the boards included in C. Accordingly, we analyse the interest of various actors before, during and after their interaction with the eWhoring community (we note some actors only interact with the eWhoring community, so they do not have interests before or after).

#Posts	#Actors	Avg. posts	%ewhor.	Before	After
≥ 1	72 982	8.8	23.3	165.3	474.2
≥ 10	13 014	37.6	22.8	142.7	449.7
≥ 50	2 146	126.9	26.0	133.8	293.8
≥ 100	815	222.4	29.1	132.8	210.1
≥ 200	263	402.3	34.9	153.6	165.7
≥ 500	46	930.8	40.6	157.4	157.8
≥ 1 000	13	1566.8	37.5	412.6	137.3

Table 8: Number of actors, mean posts made, percentage made in eWhoring and mean days posting before and after eWhoring grouped by the number of posts made in eWhoring.

6.2 Overview of actors

As shown in Table 8, we found nearly 73k actors discussing eWhoring in our dataset, i.e., they made at least one post in an eWhoring-related conversation. On average these 73k actors made 8.8 posts, and only 22% of their activity relates to eWhoring. Table 8 groups actors based on the number of eWhoring-related posts. The columns show the number of actors, the average number of posts per actor, the percentage of posts that are eWhoring-related, and the number of days they were active in the forum before and after their interaction with the eWhoring community. Figure 4 shows the Cumulative Distribution Frequency (CDF) for these metrics. Of the nearly 73k actors posting in eWhoring, around 2k made more than 50 posts, with only 46 actors making more than 500 posts.

Actors usually spend some time in the forum before interacting with eWhoring communities. Overall, actors spend around 6 months (165.3 days) before their first eWhoring-related post. Five of the most actively posting actors (those with more than 1k posts) spent more than 1 year before starting eWhoring. We observe that as users write more posts related to eWhoring, the average time spent in other sections is reduced. Also, those that make more eWhoring-related posts have a lower percentage of posts elsewhere on the forum. Thus, actors who are most active in eWhoring are more focused and have less interest in other activities.

6.3 Analysing key actors

In this section we analyse a subset of actors that are of interest due to their activities in the eWhoring community.

Definition of key actors. We focus on actors that are of interest for a variety of reasons, such as their popularity or level of reported income generated through eWhoring. We refer to these members as *key actors*. We use a rank-based selection, where a subset of top-rated users for each category are selected for analysis. We identify key actors based on the following five categories:

- Actors offering packs. In total, there are 2 523 actors who have offered packs. Of these, we select 63 actors who have shared at least 6 packs. Together, they shared a total of 554 packs (nearly 13.5% of all packs shared in *Hackforums*).
- Actors reporting substantial earnings. We rely on the self-reported proof-of-earnings to get a lower-bound estimate of the earnings made through eWhoring. Of the 661 actors posting proof-of-earnings, we include in our set of key actors the 50 highest earners. These actors claim a total of US$283k, which accounts for 55.5% of reported earnings.
- Popular actors. We identify the actors who are more popular in the community by selecting the 50 users with the highest H-index.
- Actors requesting currency exchange. We identify actors that started posting in the Currency Exchange board of *Hackforums* after first posting in eWhoring. We count the number of threads *before* and *after* their first eWhoring post. We calculate the percentage of threads made in Currency Exchange since they started eWhoring, and multiply this by the total amount of threads. Users are sorted by the resulting score, with the top 50 selected as key actors.
- Influencing actors. We leverage the social network of interactions. We calculate the eigenvector centrality, which is a metric indicating the influence of each node in the network. We select the 50 users with the highest eigenvector values.

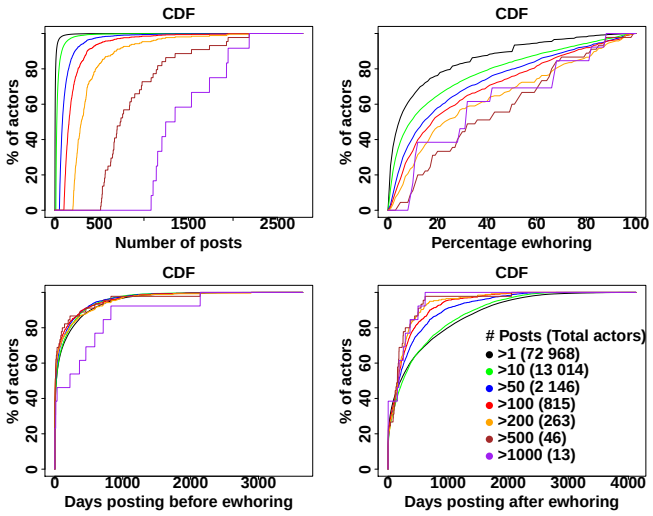

Figure 4: Cumulative frequencies of number of posts (top-left), percentage of posts in eWhoring (top-right), days posting before (bottom-left) and days posting after (bottom-right).

Selection of key actors. The intersection of the previous groups resulted in a final set of 195 key actors. Some actors belong to more than one group. Specifically, 4 actors belong to 4 groups: they are popular, influencing, offer packs and report substantial earnings. There are 16 actors from 3 groups, 14 of them are influencing and popular actors (with 9 offering packs, 4 asking for currency exchange and 1 reporting substantial earnings), one belongs to the influencing, pack offering, and earning groups; and the last one belongs to the currency exchange, popular and offering packs groups. Finally, there are 24 actors from 2 groups. The intersection between each pair of groups is shown in Table 9. The diagonal represents the number of actors unique to each category. The highest intersection is between popular and influencing actors, with 26 actors belonging to both groups. A total of 20 actors offering packs are also popular, which suggests users might offer packs for free to increase their

	Popular	Influence	Earnings	CE	Packs
Popular	11	26	10	6	20
Influence	-	19	8	4	16
Earnings	-	-	37	0	5
CE	-	-	-	44	1
Packs	-	-	-	-	40

Table 9: Number of actors selected by more than one indicator. Elements in the diagonal are unique for this indicator.

popularity. Additionally, 5 users both offer packs and report high profits. Surprisingly, none of the 50 users reporting the highest earnings are among the top 50 users asking for currency exchange. Indeed, 27 out of the 50 users reporting high earnings do not have a single thread asking for currency exchange. Either these users are exchanging profits in a different way, or they are using surrogate accounts.

Analysis of key actors. Table 10 shows the characteristics of the key actors aggregated by group. The values are the average of the actors from each group. Most actors take part in activities other than eWhoring, as the overall percentage of posts in eWhoring is low. The groups vary by number of threads in currency exchange and reported earnings. Users spent several months in the forum before starting in eWhoring. We observed some actors who spent more than a year in the forum before starting in eWhoring, becoming popular and influencing members in the community, sharing several packs and reporting thousand of US$ in earnings.

Interests of key actors. To understand the other interests of the 195 key actors, we analyse their posts made elsewhere on the forum. We divide the data into posts made by users *before* starting in eWhoring, *during* their eWhoring activities, and *after* their last eWhoring-related post. In our analysis, we removed all activity in a general board named 'The Lounge', as this is a frequently-used place for discussing miscellaneous topics. Figure 5 show the interests in the different categories of *Hackforums* by key actors before, during and after eWhoring. Initially, users are attracted by gaming and hacking boards, which is common in underground communities [24, 25]. However, after starting in eWhoring the interest for these topics decreased in favour of market-related boards. We also observe a slight increase in the *Common* category, which includes boards related to forum rules and entertainment-related topics (e.g., movies or science).

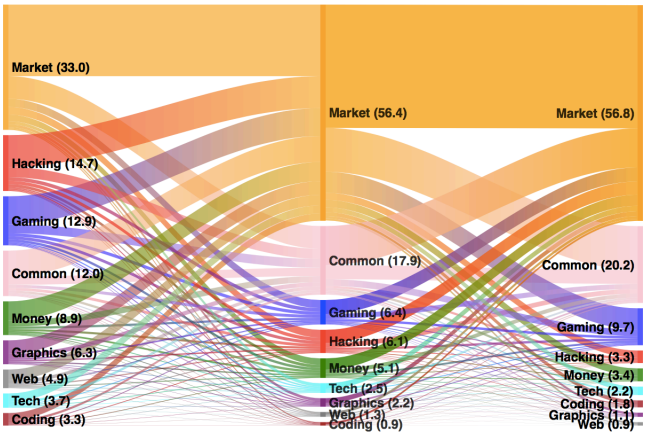

Figure 5: Evolution of interest of key actors before, during and after eWhoring. Numbers in parentheses show the percentage at each time.

7 RELATED WORK

Romance scams. eWhoring has received little research attention to date [19]. In relation to romance scams, Huang et al. [18] compared scam accounts on an online dating site, showing that these were targeted towards the demographics of intended victims. Edwards et al. [13] explore the originating country of fraudulent online dating profiles. For romance scams, the main origins are considered to be West Africa, Malaysia, or South Africa, while those in US are the prime targets. Hu et al. [17] study scam dating applications, where the platform itself is fraudulent. Payment is requested, however the profiles on the site are operated by bots.

Underground forums enabling fraud. There is a sizeable body of research into underground forums, including their role in enabling fraud and other malicious activities. Motoyama et al. [23] provides an early analysis, measuring user interactions and the types of products and services being traded in six forums. Holt et al. [16] estimate revenue generated from stolen data markets, where credit card information and account credentials are commonly traded. Most recently, Bhalerao et al. [4] uses NLP approaches to identify supply chains relating to 'hacking-for-hire' and the trade in online accounts within two forums, including *Hackforums*.

Related analysis pipelines. There is a growing interest in analysing and understanding underground or 'fringe' communities. This has lead to the development of various pipelines aimed at the automatic collection and analysis of artifacts (e.g., conversations, attachments or links) from such communities. Zannettou et al. collected nearly 9.6m posts from Twitter, 4chan and Reddit and explored dissemination of mainstream and alternative news on these communities [42]. In another paper, the authors downloaded and analysed nearly 160m images of *memes* to analyse their origin and analyse how they propagate across communities [41]. Similar to our work, authors rely on a third-party service (Know-Your-Meme) to identify meme origin. Snyder et al. proposed a pipeline to detect, collect and analyse text files related to doxing (an online abuse where offenders release private or sensitive information about victims) [33]. Authors collected 1.7m text files from three sources (4chan, Pastebin and 8ch.net), and built a classifier using NLP techniques to automatically classify text into doxing or not. As with our work, the authors relied on 'proof-of-work' files used to promote doxing-for-hire services. Data collected from underground forums and markets also play a key role to understand cybercrime communities. Soska and Christin crawled and analysed data from dark web marketplaces to analyse goods and services being traded and vendors [34]. Samtani et al. analysed source code posted in hacking forums to analyse topics of interest among hackers [30].

Group	#Posts	%eWhoring	Days before	#Amount	H	I10	I100	#Packs	#Currency Exchange
P	1 089.9	30.0	246.2	189.9	11.7	14.4	2.5	9.6	26.6
I	895.3	49.2	186.2	170.3	10.8	12.3	1.8	5.6	19.5
Hi	856.2	33.9	222.4	328.9	12.3	14.9	1.8	5.8	28.6
$	532.3	44.4	103.6	512.1	8.0	8.0	1.0	4.1	10.4
Ce	275.3	9.5	150.1	185.9	6.8	6.2	0.2	2.3	105.4
ALL	481.4	37.9	127.0	449.0	8.1	8.0	0.9	4.2	19.5

Table 10: Characteristics of key actors aggregated by groups. Values are the mean of all the actors in each group. Groups are I:influencing, P:pack providers, Hi:popular, $: reported earnings and Ce:currency exchange.

8 DISCUSSION AND CONCLUSION

This paper presents our measurements relating to eWhoring using data collected from underground forums. These forums serve as a gateway into offending by curious users aimed at easy money making methods. Packs are compiled from images and videos gathered from third-party sites (see §4.5), mostly offering adult and pornographic content (see §4.4), but also photos taken from social networks or blogs. Some of these packs were found to contain child abuse material. The proposed methodology is helpful to understand what services or sites fraudsters are being abused. For example, it could be applied by social networks to protect their users against identity theft; or from adult industry to enforce copyright.

Many of the sites where packs are distributed from have terms of service prohibiting content that infringes others' copyright or contain nudity. Indeed, we found many links where the material was removed due to infringement of these terms. As actors rely on these third-party sites for sharing material, a potential intervention approach is for image sharing and cloud storage sites to rigorously enforce their terms of service. To this end, blacklists with hashes of known images used for eWhoring, e.g. those found in packs, could be created and shared among stakeholders.

Actors post screenshots showing proof-of-earnings on the forums, with the intention of advertising service or goods or to brag about their capabilities. We analysed a set of 1.8k of these images to get an estimation of the earnings made by eWhoring. We observe actors reporting hundreds of US dollars made in a few days of eWhoring. The total, earnings reported by 661 actors are US$511k. Our analysis shows that Amazon and PayPal are the most used platforms for obtaining payment. The average amount per transaction is reported to be around US$40.

Here, payment platforms may be able to play a role in detecting and shutting down accounts used to receive payments for eWhoring. Furthermore, regulating the exchange of non-fiat currencies, such as selling gift cards for Bitcoin, may make it more difficult for criminals to monetise their activities [1].

Those legitimately involved in the adult entertainment industry potentially suffer multiple harm from eWhoring. First, their images may be stolen and used for fraud. Second, they may miss out on the revenue that has otherwise gone to a fraudster. When delivering interventions, there is the potential to cause further harm to those legitimately involved on online sex work. This should be taken into account when designing interventions, to ensure they do not have negative impacts on those that are law-abiding.

Our measurements confirm that eWhoring is a low-value but high-volume fraud. The legality of these activities is dubious, but the specifics relating to the type of crime involved will vary by jurisdiction. Until recently, it had not attracted the attention of either law enforcement or academia, probably because victims may not be aware of, or do not report, the scam. However, this activity poses various social and ethical concerns, both for offenders and victims. We have observed that users who were initially interested in gaming, coding or hacking turned into experienced eWhoring actors. These users, probably young people according to our experience reading these posts, get involved in online sexual conversations, are exposed to pornography and are at risk of downloading, storing and re-distributing illegal material. While their customers believe they are paying for encounters with attractive women, they are actually engaged in sexual activities with fraudsters, potentially underage.

ACKNOWLEDGMENTS

This work was carried out under the supervision of Dr Richard Clayton. We thank the ethics committee for help improving our design. We are also grateful to a former collaborator who decided not to proceed with this project. TinEye provided us with free access to their reverse image search API. The Internet Watch Foundation (IWF) provided us with free access to their API and information on the images we reported. We thank Pelayo Vallina-Rodriguez for his help with the domain classification study. We also thank our shepherd, Eric Wustrow, and the anonymous reviewers for their insightful comments and suggestions.

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/M020320/1], by MINECO (grant TIN2016-79095-C2-2-R), and by the Comunidad de Madrid (P2018/TCS-4566, co-financed by European Structural Funds ESF and FEDER). The opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect those of any of the funders.

REFERENCES

[1] Ross Anderson, Ilia Shumailov, Mansoor Ahmed, and Alessandro Rietmann. 2018. Bitcoin redux. In *Workshop on the Economics of Information Security (WEIS)*. University of Cambridge, Innsbruck, Austria, 33. DOI: http://dx.doi.org/10.17863/CAM.35122

[2] Ana Paula Brand ao Lopes, Sandra Eliza Fontes de Avila, Anderson Nunes Alves Peixoto, Rodrigo Silva Oliveira, and Arnaldo de Albuquerque Araújo. 2009. A Bag-of-features Approach Based on Hue-SIFT Descriptor for Nude Detection. In *Proceedings of the XVII European Signal Processing Conference (EUSIPCO)*. IEEE, Glasgow, Scotland, 1552–1556.

[3] Internet Archive. 2019. Internet Archive Wayback Machine. (2019). https://archive.org/

[4] Rasika Bhalerao, Maxwell Aliapoulios, Ilia Shumailov, Sadia Afroz, Damon McCoy, Kirill Levchenko, and Vern Paxson. 2018. Mapping the underground: Towards

automatic discovery of cybercrime supply chains. *arXiv preprint arXiv:1812.00381* (2018), 16.
[5] blackhatworld.com. 2019. Terms of Service and Rules. BlackHatWorld. (March 2019). https://perma.cc/S9C7-JQ4T
[6] British Society of Criminology. 2015. Statement of Ethics. (2015). http://www.britsoccrim.org/ethics/
[7] Mark Button and Cassandra Cross. 2017. *The 'fraudogenic' consequences of the Internet revolution.* Routledge, New York, 78.
[8] Andrew Caines, Sergio Pastrana, Alice Hutchings, and Paula J. Buttery. 2018. Automatically identifying the function and intent of posts in underground forums. *Crime Science* 7, 1 (2018), 19.
[9] Bill Chu, Thomas J. Holt, and Gail Joon Ahn. 2010. *Examining the creation, distribution and function of malware on-line: Technical Report for National Institute of Justice.* Technical Report. U.S. Department of Justice. https://www.ncjrs.gov/pdffiles1/nij/grants/230111.pdf
[10] Jonathan Clough. 2015. *Principles of Cybercrime.* Cambridge University Press, Cambridge, UK.
[11] David Dittrich and Erin Kenneally. 2012. *The Menlo Report: Ethical principles guiding information and communication technology research.* Technical Report. US Department of Homeland Security.
[12] Greg Durrett, Jonathan K. Kummerfeld, Taylor Berg-Kirkpatrick, Rebecca Portnoff, Sadia Afroz, Damon McCoy, Kirill Levchenko, and Vern Paxson. 2017. Identifying products in online cybercrime marketplaces: A dataset for fine-grained domain adaptation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics, Copenhagen, Denmark, 2598–2607.
[13] Matthew Edwards, Guillermo Suarez-Tangil, Claudia Peersman, Gianluca Stringhini, Awais Rashid, and Monica Whitty. 2018. The geography of online dating fraud. In *Workshop on Technology and Consumer Protection (ConPro).* IEEE, San Francisco, 7.
[14] Jason Franklin, Vern Paxson, Adrian Perrig, and Stefan Savage. 2007. An inquiry into the nature and causes of the wealth of Internet miscreants. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS).* ACM, Alexandria, VA, USA, 375–388. DOI: http://dx.doi.org/10.1145/1315245.1315292
[15] Thomas J. Holt. 2013. Examining the forces shaping cybercrime markets online. *Social Science Computer Review* 31, 2 (2013), 165–177.
[16] Thomas J Holt, Olga Smirnova, and Yi Ting Chua. 2016. Exploring and estimating the revenues and profits of participants in stolen data markets. *Deviant Behavior* 37, 4 (2016), 353–367.
[17] Yangyu Hu, Haoyu Wang, Yajin Zhou, Yao Guo, Li Li, Bingxuan Luo, and Fangren Xu. 2019. Dating with scambots: Understanding the ecosystem of fraudulent dating applications. *IEEE Transactions on Dependable and Secure Computing* (2019), 18. DOI: http://dx.doi.org/10.1109/TDSC.2019.2908939
[18] JingMin Huang, Gianluca Stringhini, and Peng Yong. 2015. Quit playing games with my heart: Understanding online dating scams. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA).* Springer, Milan, Italy, 216–236.
[19] Alice Hutchings and Sergio Pastrana. 2019. Understanding E-Whoring. In *Proceedings of the IEEE European Symposium on Security and Privacy (EuroSP).* IEEE, Stockholm, Sweden, 14.
[20] Tom Hymes. 2013. Raging Stallion faces cloud service Oron in Ninth Circuit appeal. (2013). https://avn.com/business/articles/legal/raging-stallion-faces-cloud-service-oron-in-ninth-circuit-appeal-517690.html
[21] Jay Mahadeokar and Gerry Pesavento. 2016. Open sourcing a deep learning solution for detecting NSFW images. (September 2016). https://perma.cc/7C88-Y5BK
[22] Microsoft. 2009. PhotoDNA. (2009). https://www.microsoft.com/en-us/PhotoDNA
[23] Marti Motoyama, Damon McCoy, Kirill Levchenko, Stefan Savage, and Geoffrey M. Voelker. 2011. An analysis of underground forums. In *Proceedings of the Internet Measurement Conference (IMC).* ACM, Berlin, Germany, 71–80.
[24] National Crime Agency. 2017. Pathways into Cyber Crime. (2017). https://perma.cc/897P-GZ3R
[25] Sergio Pastrana, Alice Hutchings, Andrew Caines, and Paula Buttery. 2018. Characterizing Eve: Analysing cybercrime actors in a large underground forum. In *Research in Attacks, Intrusions, and Defenses (RAID).* Springer, Heraklion, Crete, Greece, 207–227.
[26] Sergio Pastrana and Guillermo Suarez-Tangil. 2019. A first look at the crypto-mining malware ecosystem: A decade of unrestricted wealth. In *Proceedings of the Internet Measurement Conference (IMC).* ACM, Amsterdam, Netherlands.
[27] Sergio Pastrana, Daniel R. Thomas, Alice Hutchings, and Richard Clayton. 2018. CrimeBB: Enabling cybercrime research on underground forums at scale. In *Proceedings of the World Wide Web Conference (WWW).* International World Wide Web Conferences Steering Committee, Lyon, France, 1845–1854. DOI: http://dx.doi.org/10.1145/3178876.3186178
[28] Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, Christian Kreibich, and Phillipa Gill. 2018. Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem. In *25th Annual Network and Distributed System Security Symposium (NDSS).* Internet Society, San Diego, California, USA, 15. DOI: http://dx.doi.org/10.14722/ndss.2018.23353
[29] Aunshul Rege. 2009. What's love got to do with it? Exploring online dating scams and identity fraud. *International Journal of Cyber Criminology* 3, 2 (2009), 494–512.
[30] Sagar Samtani, Ryan Chinn, and Hsinchun Chen. 2015. Exploring hacker assets in underground forums. In *International Conference on Intelligence and Security Informatics (ISI).* IEEE, Baltimore, MD, USA, 31–36. DOI: http://dx.doi.org/10.1109/ISI.2015.7165935
[31] Ray Smith. 2007. An overview of the Tesseract OCR engine. In *9th International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 2. IEEE, Parana, Brazil, 629–633.
[32] Ray Smith and Zdenko Podobny. 2018. Tesseract open source OCR engine (main repository). (October 2018). https://github.com/tesseract-ocr/tesseract
[33] Peter Snyder, Periwinkle Doerfler, Chris Kanich, and Damon McCoy. 2017. Fifteen minutes of unwanted fame: Detecting and characterizing doxing. In *Proceedings of the Internet Measurement Conference (IMC).* ACM, London, UK, 432–444. DOI: http://dx.doi.org/10.1145/3131365.3131385
[34] Kyle Soska and Nicolas Christin. 2015. Measuring the longitudinal evolution of the online anonymous marketplace ecosystem.. In *Proceedings of the USENIX Security Symposium.* USENIX, Washington, D.C., USA, 33–48.
[35] tesseractocr. 2019. GUIs and Other Projects using Tesseract OCR. (2019). https://github.com/tesseract-ocr/tesseract/wiki/User-Projects-\T1\textendash-3rdParty Accessed September 2019.
[36] TinEye. 2008. TinEye. Reverse Image Search. (2008). https://www.tineye.com/
[37] Ramarathnam Venkatesan, S-M Koon, Mariusz H. Jakubowski, and Pierre Moulin. 2000. Robust image hashing. In *Proceedings International Conference on Image Processing (ICIP)*, Vol. 3. IEEE, Vancouver, BC, Canada, 664–666.
[38] Monica T. Whitty and Tom Buchanan. 2012. The online romance scam: A serious cybercrime. *CyberPsychology, Behavior, and Social Networking* 15, 3 (2012), 181–183.
[39] Michael Yip, Craig Webber, and Nigel Shadbolt. 2013. Trust among cybercriminals? Carding forums, uncertainty and implications for policing. *Policing and Society* 23, 4 (2013), 516–539.
[40] Kan Yuan, Di Tang, Xiaojing Liao, XiaoFeng Wang, Xuan Feng, Yi Chen, Menghan Sun, Haoran Lu, and Kehuan Zhang. 2019. Stealthy Porn: Understanding Real-World Adversarial Images for Illicit Online Promotion. In *IEEE Symposium on Security and Privacy (SP).* IEEE Computer Society, San Fransisco, CA, USA, 547–561. DOI: http://dx.doi.org/10.1109/SP.2019.00032
[41] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference (IMC).* ACM, Boston, MA, USA, 188–202. DOI: http://dx.doi.org/10.1145/3278532.3278550
[42] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the Internet Measurement Conference (IMC).* ACM, London, UK, 405–417. DOI: http://dx.doi.org/10.1145/3131365.3131390
[43] Xiong Zhang, Alex Tsang, Wei T. Yue, and Michael Chau. 2015. The classification of hackers by knowledge exchange behaviors. *Information Systems Frontiers* 17, 6 (01 Dec 2015), 1239–1251. DOI: http://dx.doi.org/10.1007/s10796-015-9567-0

APPENDIX: ETHICAL ISSUES

This research involved a substantial number of ethical issues. To address them,we used the guiding principles contained in *The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research* [11] and the British Society of Criminology's *Statement of Ethics* [6]. We also complied with the Cambridge Cybercrime Centre's data sharing agreements, which sets out how the data may be used and handled. Additionally, the design of our research was approved by the Research Ethics Board (REB) of our institution. We next describe in detail our approach with respect to these frameworks.

P1 Respect for Persons

Stakeholder identification. The dataset we used to conduct this research contains the digital identifiers used by actors involved

in eWhoring forums, as well as their communications and other relationships that can be derived from such identification. Even if these identifiers are merely nicknames, this could imply a risk to those actors. We took precautions to avoid exposing the identity of these research subjects. In accordance with the British Society of Criminology's *Statement of Ethics*, our research analyses only collective behaviour, rather than individuals. We were also careful with our presentation of results to reduce the likelihood they could be linked back to the forum accounts, potentially leaking their identity. The dataset we were provided with[8] was collected from public sources and already contained the identifiers. The identifiers are hard to fully remove as (variations on) the identifiers are used in the text of posts. Hence, removing the identifiers would make the analysis more difficult but not reduce risk to forum users as we do not publish anything that can be linked to those identifiers.

Informed consent. Our analysis involves subjects whom it is impossible to obtain informed consent from, since this did not took place at collection time when the dataset was created. This, however, fits the case discussed in *The Menlo Report* in which gathering informed consent is impracticable:

> *'Because of the difficulty in identifying all individuals from whom consent should be sought or in practicably obtaining consent, researchers or REBs may frequently conclude that seeking a waiver of informed consent or waiver of documentation of informed consent are the only options'* [11].

We therefore resorted to a waiver of informed consent issued by our REB. This approach also complies with the British Society of Criminology's *Statement of Ethics* [6], as the dataset is collected from online communities where the data are publicly available, and, as mentioned above, is used to analyse collective rather than individual behaviour.

P2 Beneficence

Identification of Potential Benefits and Harms. Our work involves societal benefits as it attempts to provide an improved understanding of a phenomenon with potential legal implications and, by extension, an advancement of existing knowledge of the cybercrime ecosystem. Furthermore, our analysis pipeline could find application in other cybercrime areas, thus fostering and facilitating additional research in the field. However, conducting this research involves the analysis of images containing sexual or nude content. The viewing of such images poses a risk to the well-being of researchers. Additionally since these images come from unknown parties engaged in cybercrime, there is a risk that there may be indecent images of children included within the images that we would collect. Possession of such images is a strict liability offence in the jurisdictions where this research was conducted. Hence, the possession of such images presents a legal risk.

Balancing Risks and Benefits. Our assessment of risks and benefits concluded that there is a reasonable balance that justifies this work if a number of design precautions were implemented to mitigate or minimise risks. We detail them next.

Mitigation of Realised Harms. Our analysis pipeline (§ 4) was purposely designed to address the risks mentioned above, thus limiting the ethical and legal concerns. Classification and analysis of sensitive material (particularly images) is done semi-automatically, using established methods and tools to avoid researchers' exposure to potentially perturbing material.

Our research design was based on the assumption that finding indecent images of children was very unlikely. Nonetheless, we discussed this research with the IWF in advance in order to reduce the likelihood of prosecution, and guidelines were established in the event that images were classified as child exploitation material, namely immediately reporting the location of the images to the IWF, and deleting them from our servers. These guidelines were established as a precaution with the assistance of the REB. We originally thought it unlikely that the dataset would contain child abuse material, and were surprised when matches were identified. We reiterate that these images were never viewed by any of the researchers, and our guidelines were complied with. Given our experience, future research in this area should assume a greater likelihood of finding indecent images of children, and so would need to consider whether our design is adequate.

P3 Justice

Our study complies with the principles of fairness and equity in the selection of research subjects. We do not arbitrarily target or exclude users based on any attribute other than an active participation in eWhoring forums. We do not know the representativeness of that user base in terms of religion, political affiliation, sexual orientation, health, age, technical competency, national origin, race, or socioeconomic status, as suggested in the Menlo Report [11].

P4 Respect for Law and Public Interest

Compliance. Our prior belief was that the risk of obtaining indecent images of children was very low, as it is against the rules of the forums we were looking at and we expected moderators would remove any such content. However, as noted above, we took precautions against this eventuality and, as discussed in (§4.3), we were surprised to find that a small proportion of the images were indecent images of children. Those were reported using appropriate channels (the INHOPE hotline operator for the jurisdiction in which the work was conducted).

Transparency and Accountability. In our reporting of this research, we attempted to follow the principles of transparency and accountability. This encompasses our data sources, research methodology and tools used (including known or identified limitations), and the main results obtained. Our tools, data, and acquired know-how remains at the disposal of the community for any enquiry that requires further clarifications. No data or insights that could help to identify users will be shared.

[8] This dataset has also been provided to 19 other research groups internationally.