Getting Bored of Cyberwar: Exploring the Role of Civilian Hacktivists in the Russia-Ukraine Conflict

Anh V. Vu†, Daniel R. Thomas‡, Ben Collier‡, Alice Hutchings‡, Richard Clayton†, Ross Anderson†‡

†University of Cambridge, ‡University of Strathclyde, †University of Edinburgh

ABSTRACT

There has been substantial commentary on the role of cyberattacks and civilian hacktivists in the Russia-Ukraine conflict. Drawing on a range of data sources, we argue that the widely-held narrative of a significant cyberwar fought by committed civilians and volunteer ‘hacktivists’ linked to cybercrime groups has likely been overhyped. We collected 358k web defacement attacks, 1.7M reflected DDoS attacks, and 441 announcements (with 58k replies) of a volunteer hacking discussion group for two months before and four months after the invasion. To enrich our quantitative understanding, we conducted interviews with individuals who were active in defacing Russian and Ukrainian websites. Our findings indicate that the conflict briefly but significantly caught the attention of the low-level cybercrime community, with notable increases in both defacement and DDoS attacks targeting Russia and Ukraine. However, the role of these players in the so-called cyberwarfare is minor, and they do not resemble the ‘hacktivists’ imagined in popular criminological accounts. Initial waves of interest led to more attackers participating in defacement campaigns, but rather than targeting critical infrastructure, there were mass attacks against random websites within ‘.ru’ and ‘.ua’. We find little evidence of high-profile actions of the kind hypothesised by the prevalent narrative. The much-vaunted role of the IT Army of Ukraine co-ordination group is mixed; their promoted targets were seldom defaced although sometimes subjected to DDoS attacks. Our main finding is that there was a clear loss of interest in carrying out defacement and DDoS attacks after just a few weeks. Contrary to the prediction of some commentators, the involvement of civilian hacktivists from low-level crime groups in the conflict appears to have been minor, short-lived, and fleeting.

CCS CONCEPTS

• Social and professional topics → Computer crime; • Applied computing → Cyberwarfare; • Security and privacy → Social aspects of security and privacy; • Mathematics of computing → Time series analysis; • Networks → Denial-of-service attacks.

KEYWORDS

Web defacement; DDoS attacks; cybercrime; civilian participation; volunteer hacktivists; IT Army of Ukraine; Russia-Ukraine conflict.

1 INTRODUCTION

Researchers, politicians, and journalists have long been fascinated by ‘cyberwar’ – the spectre of armed conflict between nations spilling over into attacks conducted over the Internet [77]. ‘Colder’ forms of inter-state conflict are characterised by espionage and intelligence gathering, which may facilitate the degradation of online systems once hostilities commence [27]. Alongside this there has been a thirty-year history of speculation around how the tools and techniques of the cybercrime underground – Distributed Denial of Service (DDoS) attacks, disruption and compromise of services, web defacements, and similar techniques – might allow civilians to play a role in a ‘hot’ war between developed nations [5]. Much of this speculation, drawing from criminological models of low-level underground cybercrime groups and on links between this underground and well-organised ‘hacktivist’ movements, has argued that these groups would play a crucial role, making the future of war hybrid, chaotic, and unpredictable [95]. The 2022 invasion of Ukraine provides an opportunity to assess what has happened in practice.

Russia and Ukraine have a long history of electronic information warfare [45] and are among the most active cybercrime hubs [57]. When Russia invaded Ukraine on 24th February 2022, war-related attacks hitting the two countries were regularly reported [15, 23, 99]. There was an expectation of civilians and volunteers getting involved in a ‘cyberwar’; a popular narrative is that their engagement could be a game changer and could undermine Russia’s war [73]. Some commentators have predicted it will be the “first full-scale cyberwar” [24], its effects will last for decades [81], and youngsters would be drawn into cyberattacks by joining the ‘IT Army of Ukraine’ – a group backed by the Ukrainian state to co-ordinate volunteer hacktivists and civilians to help disrupt Russian assets [31]. The security industry has suggested that there would be a “real cyber war”, predicting hacktivist attacks on Russia would escalate further throughout 2022 [2]. These narratives regularly appear in the press and play a role in shaping domestic policy responses to cybercrime. Although less likely to grab headlines, a contrary narrative around ‘overhyped cyberwar’ suggests that cyber operations have been too slow and too insignificant in the ongoing hybrid conflict [59]. GCHQ commented the cyber conflict has not yet materialised [28] and point to the resilience of Ukraine’s defences [87].

We explore the role civilian hacktivists have played in the conflict. Our study starts with quantitative longitudinal measurements on datasets associated with the low-level cybercrime activities, including web defacement and reflected DDoS attacks (§4). The findings are enhanced with an analysis of discussions from a pro-Ukraine volunteer group promoting cyberattacks against Russia (§5), then strengthened by qualitatively interviewing individuals responsible for a significant number of defacements targeting either Russia or Ukraine (§6). This is the first academic work to explore the engagement of civilian hacktivists in a hybrid war, in which both sides have substantial IT infrastructure, a thriving digital underground crime ecosystem, and significant access to offensive capacities.

Our first impression is that there was an immediate increase in the proportion of defacements targeting Russia after the invasion (Figure 1). However, it appears that alarmist predictions about a ‘cyberwar’ producing organised, motivated, and technically skilled civilian and volunteer ‘hacktivists’ have not come to pass. Instead, we find that most budding cyberwarriors used trivial attacks to take down minor targets and largely got bored after a few weeks. Our
study was approved by our institutional Ethics Review Board, allowing us to collect, analyse the data, and conduct interviews while minimising risks to participants (see Appendix A). Our data and scripts are available to academic researchers (see Appendix B). A dashboard to monitor country-based defacement and UDP amplification DDoS attacks over time is available at http://cyberstrikes.live/

2 BACKGROUND

The development of Internet measurement scholarship in recent years has contributed to, and in some cases challenged, criminological understanding of the drivers of cybercrime. Measurement studies of the impact of externalities on cybercrime and policing have been explored; police interventions in DDoS-for-hire markets have shown distinct but temporary effects [22, 51]; the pandemic has affected illicit trading activities [94], scam and fraud [14, 18, 49], online victimisation [35], novel cybercrime vectors [21], listings on darkweb markets [8–10]; and cybercrimes have been connected to increased long-term psychological and financial consequences for victims [62]. Yet, there has been limited empirical academic studies looking at the relationship between armed conflicts and cybercrime. Some have been conducted in the private sector [64, 65]; the incident response team of a Czech university has reported negligible impact on their network after hundreds of their users launched DDoS attacks against Russia for a week after the invasion [40].

Information warfare has long been a routine part of ‘hybrid’ modern conflicts, especially around controlling communications [37, 55]. Cryptography has thousands of years of history and its contemporary usage in the NATO context includes a range of offensive cyber operations [91]. The enemy’s capability to spread news and propaganda can be impacted by targeting crucial sites, public services, broadcast and telecommunication infrastructure. Russia has blocked news and anti-war domains since the conflict started [75, 84], and has lost access to foreign service providers [47] and websites [75]. Ukrainian users experienced degraded network performance [44], while pro-Ukrainian supporters have tried unconventional channels such as online reviews to bypass the censorship [63]. Attacks are not just online – Russian missiles hit TV towers in Kyiv in early March 2022 [34]. Censorship is often used during wartime [74]; governments block access to global services, especially social networks, news, and media platforms to suppress unwanted narratives.

An association between kinetic warfare and ‘nationalistic’ cyberattacks has been reported with various information operations. Ukrainian firms got hit by data wipers such as CaddyWiper and NotPetya [1, 68], DDoS attacks [12, 83] and phishing campaigns [53], while Ukraine supporters have used spam senders to distribute propaganda in Russia [89] and have stolen cryptocurrency from Russian wallets [93]. Ukrainian universities were hacked [99], the Ukrainian electricity grid was hit by Industroyer2 [36], and the Ukrainian satellite Internet was downed by Russia [76]. Attackers identifying themselves under the banner of the Anonymous movement declared a ‘cyberwar’ on Russia [61] with attacks against Russian Ministry of Defence databases [54] and state TV channels [66]. Russian systems intermittently received cyberattacks instigated by volunteer hacktivists from the IT Army of Ukraine [16, 73].

Government-backed phishing campaigns [29] and routine acquisitive crime have continued. Yet, those crimes are reportedly state-sponsored [97], and data about state-level attacks is extremely hard for academics to collect. We thus do not cover financial crime, fraud, data leaks, phishing, malware or ransomware. Crucially, we do not study the impact of ‘leaking’ campaigns by established non-criminal hacktivists, as these do not fit the model of a galvanised ‘big fish’ behind significant real-world attacks may not list themselves in public (which is also observable from our data and interviews). We are particularly interested in the engagement of non-governmental activity contributed by many ‘small fish’ such as civilian hacktivists, instead of state actors. We focus on the hypothetical ‘volunteer army’, particularly their offensive tools for launching attacks.

One type of attacks linked with the low-level ‘script kiddie’ is web defacement [80]. It accounted for around 20% of online attacks [69] and is often organised into discrete campaigns [58]. Attackers (or defacers) gain unauthorised access using off-the-shelf tools and simple exploits, then alter sites’ appearance by their own messages.
Another simple type of attack is amplified DDoS, which floods targets with unwanted traffic, thus prevents them from servicing expected system, its IP address and location, and a snapshot of the defaced page is taken (often consisting of the defacer’s messages, promised system, its IP address and location, and a snapshot of the defaced page is taken (often consisting of the defacer’s messages). Launching them with ready-made tools is straightforward for civilian hacktivists without much technical expertise. They can be executed quickly, at scale, and have instant, noticeable effects such as altering targets’ appearance, taunting opponents with compromised websites, or making it inaccessible. During wartime, the need to rapidly disseminate political messages and propaganda makes these types of attack attractive.

3 METHODS AND DATASETS

Our mixed-methods approach enriches quantitative evidence with complementary qualitative research. First, we drew insights and interpretations based on three quantitative datasets collected regularly and separately: defacements, amplified DDoS attacks, and discussions from the IT Army of Ukraine’s Telegram channel. Our data covers 2 months before and 4 months after the invasion, spanning 1 January to 30 June 2022. Second, we interviewed involved attackers – focusing on those actively defacing Russian or Ukrainian websites – to illuminate our quantitative findings and develop a better understanding of how the conflict influenced their activities.

3.1 Web Defacement Attacks

The largest and most popular active defacement archive is Zone-H, which has been online for over two decades since March 2002. It provides cybersecurity news and self-reported defacements along with hacking content [52], holding over 15M reports (around 2 000 per day). The others are more recent, functioning similarly: OwnzYou (since January 2021), Zone-Xsec (since May 2020), Haxor-ID (since November 2019), and Defacer-Pro (since June 2021) with around 140k, 340k, 110k, and 92k reports in total, respectively.\(^1\) While not all compromised websites will be reported, measuring trends from the widely used archives is still informative. We identify the country of defaced websites based on ccTLD, IP geolocation, and geolocation of the AS hosting that IP, excluding CDNs (see Appendix C). The defacement submission process is mostly automated. Users specify (if any) a ‘notifier’, team information, defaced URL, type of vulnerability, and hacking incentives. New reports are kept away from the main dashboard as ‘on hold’ until being verified either by staff or bots. At that point, a record is made with details of the compromised system, its IP address and location, and a snapshot of the defaced page is taken (often consisting of the defacer’s messages, which may include political and ideological propaganda [6]). Although ‘notifier’ can be arbitrarily entered, defacers are incentivised to use a consistent handle to cultivate fame and reputation. We thus consider ‘notifier’ to be reliable enough to differentiate between

\(^1\) Our participants confirm Zone-H is the most trusted, while the others are less reputable. Smaller archives were historically active [58], but either vanished (Hack Mirrors and Mirrors Zone) or have changed to host different content (Hack-CN and MrDeface).
defacers. We also consider snapshots of defaced websites, including messages left, to be reliable, as they are captured at reporting time.\(^2\)

We use web scrapers to collect defacements (see Appendix D). The process is not straightforward as we need to ensure data completeness and bypass blocking challenges such as Captcha and IP blacklisting. Our collected data is detailed in Table 1.\(^3\) Further steps are performed to enhance data reliability. First, many on-hold submissions are valid but were never verified; defacers in our later interviews confirmed the necessity to look at both the archives and on-hold submissions. We thus perform a semi-automated validation of on-hold reports using the messages left on the defaced pages (see Appendix E). Second, submissions may be reported to multiple archives to broaden their visibility. We de-duplicate across and within archives by hashig their content (see Appendix F). Third, as ‘notifier’ can be arbitrary, typos can give a single attacker multiple identities. So we correct typos by comparing handles’ similarity, and checking against messages left on the defaced pages (see Appendix F). This handle unification is done across all archives.

In total, 137 339 reports were verified by the archives, 97 652 were automatically validated by us and a further 39 972 were validated semi-automatically. We consider 42 086 to be invalid or not decidable due to lack of evidence. Zone-Xsec, Haxor-ID, and Defacer-Pro have high validity rates (99.80%, 99.92%, and 95.36%, respectively), suggesting their automatic sanitisation works well at reducing invalid submissions. The validity rates of Zone-H and OwnzYou are lower at 87.32% and 70.59%, respectively. As a cross-check, Zone-Xsec and Haxor-ID have low rates of defacers submitting at least one invalid record (only 4.43% and 3.37%), and high rate of defacers with at least one valid submission (99.82% and 99.55%, respectively). Zone-H, OwnzYou and Defacer-Pro have high rates of at least one invalid submission per defacer (41.77%, 70.51% and 29.58%, respectively), but Defacer-Pro has a higher rate of defacers submitting valid reports, 97.79% compared to 82.01% of Zone-H and 54.00% of OwnzYou. These figures once more suggest the automatic validity sanitisation is effective. Duplicate submissions are removed: 40 799 (11.00%) are merged across all archives; the remaining are 317 049 with 274 963 valid defacements (86.73%) are used for our analysis, in which 4 347 defacer handles are also unified to 3 454. Notably, two handles of a defacer who was very actively targeting Russia during our study period were unified. Our defacement collection has only around 2 000 reports per day, but collates all reports on the most trusted archives used by the defacement community.

### 3.2 UDP Amplification DDoS Attacks

We use 1.7M DDoS attack records gathered by a honeypot network deployed worldwide since 2014, which emulates UDP protocols vulnerable to reflected UDP amplification DDoS attacks and are discovered by people scanning for reflectors [88]. The honeypots record scanning traffic and subsequent attacks, but avoid forwarding amplified traffic toward victims. The nature of these attacks means that only victim information is recorded and the attack origin is not available. This dataset has previously been used to evaluate the success of various interventions against booter operators [22], and has good coverage for the protocols emulated, though there is a risk of honeypots being systematically avoided by attackers. It only covers reflected attacks, while some may have been directly sent from botnets, compromised servers and misconfigured proxies.

In 2022, the median number of honeypots contributing data was 50, 95% CI [34, 51]. A flow of packets is considered to be an attack if any sensor observes at least 5 packets intended for the same victim IP or IP prefix, and the attack is deemed to last from the first packet until the last packet preceding a 15 minute period without further packets. In 2022, the observed median number of attacks per week was 35 000, 95% CI [11 900, 271 000] and on IP prefixes of 438, 95% CI [0, 3 480]; the median duration of attacks was 1.53 minutes, while the maximum was 11 300 minutes. The country of DDoS victims is identified based on IP geolocation and geolocation of the AS hosting that IP, excluding CDNs (see Appendix C).

### 3.3 Volunteer Hacking Discussions

Recent years have seen increasing use of messaging apps such as Telegram and Wickr to co-ordinate cybercrime, with underground communities supplementing traditional forum-based modes of organisation with a range of other channels, particularly for more sensitive conversations. A couple of days after the invasion, the Ukrainian government called on pro-Ukraine ‘hackers’ to join the ‘IT Army of Ukraine’ to support the war effort [31, 33].\(^4\) The most tangible outcome is a public Telegram channel mainly used to recruit and encourage those with relevant skills – whether employed in legal cybersecurity work or in more criminal ventures – to spread pro-Ukraine news, propaganda, and co-ordinate disruptive efforts against Russian infrastructure. They use a range of attack vectors (but mostly DDoS) to disrupt communication and financial systems by hitting banks, businesses, government, and logistics [16].

The Telegram channel, attracting more than 200k subscribers, provides access to guides and tools for launching attacks.\(^5\) Most days there were announcements in Ukrainian and English advertising new lists of Russian targets (hostnames and IP addresses). This ‘cyber army’ has claimed that ordinary Russians have seen impacts when they hit banks, exchanges [11], and cinemas [72]. They claimed to have been significant enough that many important Russian sites (e.g., SberBank) have taken additional defensive measures (e.g., restricting access to Russian IP addresses only, using third-party DDoS protections) after being attacked. We confirmed with a Ukrainian government source that the Telegram channel we analysed was the primary one used for communication amongst Ukrainian civilians, with messages being forwarded to the other unofficial satellite groups which have far fewer subscribers.

We use Teletthon to collect 441 announcements with 57 757 replies and 900k emoji reactions posted on the official IT Army of Ukraine’s Telegram channel from its creation until 30 June 2022. The tools

---

\(^2\) Captured snapshots may contain hidden messages with identical font colours as the background, but they are still detectable and extractable by analysing the HTML.

\(^3\) An active archive Defacer-ID (since February 2016) is excluded because (1) the valid submission volume during the period is small (less than 27K); (2) unclear staff verification, no validity sanitisation on submission, no validity signal in defaced pages (in fact, over half of these have been deemed invalid by the archive); (3) snapshots of defaced pages and defacers’ messages are missing; and (4) victim geolocation is mostly lacking; determining it after the fact is problematic as sites could have been relocated.

\(^4\) IT Army of Ukraine: https://itarmy.com.ua/

\(^5\) DDoS tools for individuals: ‘Death by 1000 needles’ tool to quickly fetch daily targets; tool for granting access to individuals’ cloud resources for later coordinated attacks.
interact with official Telegram APIs, so they can fully capture messages and metadata. We use this dataset to track the choice of targets and cybercrime activities of volunteers as the conflict progresses.

### 3.4 Interviewing Website Defacers

As a handful of attackers were responsible for the majority of defacements (see more in 4.3), we interviewed two participants from two teams responsible for defacing many Russian and Ukrainian websites when the war began. The interviews were designed to gain a qualitative understanding of defacement activities and motivations, how they shifted over time (particularly during the war) and the impact of the conflict on the choice of targets. We also approached other prolific defacers who did not target Russia or Ukraine during the period, but they either did not reply or declined to be interviewed — “They do not like to be famous or share their secrets; they are very silent”, said one participant. We thus extended our interviews to ask further questions about the wider community’s opinions and war-related behaviours, as many defacers are active in a private chat group, and claim to know each other well.

Participants’ responses cannot be generalised, but they help inform findings derived from our quantitative measurements as they represent insights from individuals directly involved in the data generation. They are substantially useful in interpreting the movements and associations between and within our measurement datasets. We present these qualitative interviews not as a robust sociological study of these communities, but rather as a way of gaining insight into our quantitative collections and analysis, in the tradition of previous social data science and digital phenomenological studies [17]. These use partial qualitative collections to enhance the analysis of quantitative evidence generated in complex environments with a range of potential determining factors [7]. As quantitative data can be explained in different ways, the incorporated qualitative interpretation strengthens our conclusions. The interview participants have a clear incentive to overstate their own skill and importance, we thus have taken care to present their statements objectively.

We recruited our participants by reaching out to social media accounts of teams linked with defaced websites. The first to reply referred us to the second participant. Our participants were members of two 'black-hat' hacking teams (as they claimed), each supporting different sides (Russia or Ukraine). One identified themselves as the ‘team owner’, while the other claimed to be a key official member responsible for all of his team’s attacks. To establish trust, we explicitly let them know our identity as academic researchers, the reasons for our interest, and we did not ask questions that could potentially de-anonymise them. We used a semi-structured interview schedule, which included warm-up questions, followed by a lengthy interview in which follow-up questions are adapted flexibly based on previous responses (see Appendix G).

We followed a strict protocol for the safety and privacy of participants and researchers. First, we obtained ethics approval from our institution (Appendix A). We then contacted participants and invited them to the interview. An informed consent package introducing our research project was delivered prior to the interview, which clearly explained the topics to be covered and potential risks to participants. At the start of the interview, participants consented to be interviewed and for their team information to be explicitly disclosed in this paper. Interviews were conducted by text chat on Telegram only (no audio or video recording). Participants could refuse to answer any question. The interview responses were synthesised, then all chats were permanently destroyed. Participants were provided with a copy of our synthesised findings to check for misunderstandings or potentially identifying information.

The first participant is part of a Kurdish competitive defacement team founded in 2019. They said that they initially attacked government websites for political reasons, but then expanded to include businesses, banks and other government sites. The team has links with Kurdish nationalism, when Turkey and Iran were attacking Kurds, the team claimed to have hit many Turkish and Iranian targets and to have left messages to support their cause. This team is among the top 10 active defacers since January 2022, and they were the fourth most active team targeting Russia when the war began. The second participant claims to be a member of the first Brazilian group acting outside of Brazil (founded in 2019, now disbanded). They said that they initially only attacked Brazilian targets, but then escalated to other countries, mostly focusing on government and educational sites. They have not been among the top ten defacers since January 2022, but when the war started they were the second most active team targeting Ukrainian websites.

### 4 THE EVIDENCE OF CYBERATTACKS

Having initially detected the increase in defacements targeting Russia (Figure 1), we started to measure the shift of cyberattacks in detail, both on the Russia-Ukraine and global scales. This section discusses the quantitative evidence we have seen, and what we have learnt about involved actors as the war progressed. Only validated defacement attacks are used; timestamps are normalised to UTC.

#### 4.1 Attacks Targeting Russia and Ukraine

Figure 2 shows defacement and DDoS attacks hitting Russia and Ukraine over the period. Figure 3 shows the hour by hour changes for the most active four-week period from 17 February for defacements, and from 24 February for DDoS attacks where the activity was almost after 7AM (we count both DDoS victims and DDoS attacks as one victim can be associated with many attacks). The same pattern arises for both types of attack: Russia was the first to be attacked at scale, followed by Ukraine a few days later. The key observation is that in both countries, the outbreak of attacks was fairly short-lived: it returned to pre-war levels after a few weeks.

The number of defacements targeting Russia peaked on the invasion day at 209 (14.48% of all defacements on that day, while it was 0.60% the day before): the first big wave occurred at around 10AM (7 hours after the 3AM invasion) with 178 attacks caused by a single defacer, followed by smaller waves on the same day. Two follow-up waves occurred over the next two days at 1PM on 25 February and 9AM on 26 February with 43 and 109 attacks, respectively. The number of defacers targeting Russia peaked two days later: while only 11 defacers were responsible for the peak on 24 February, it was 22 on 26 February. As to defacements on Ukraine, there were no notable changes on the invasion day, but a peak of 69 attacks occurred two days later (6.30% of all defacements on that day, while it was 0.47% the day before). The largest wave was at around 7PM on 26 February with 50 attacks, followed by medium
waves at 5PM on 27 February (26 attacks) and 10PM on 3 March (29 attacks). The number of defacers targeting Ukraine peaked on 27 February (one day after the largest wave) with 9 defacers.

These results show waves of attacks against both countries directly after the invasion. The number of defacers involved was quite small, but nevertheless, at least momentarily they turned from indiscriminate targets (seeking out common vulnerabilities on websites), to more targeted attacks. From our interviews, there was a 'call for hacking' within the defacers’ community after the initial flurry of activity; many were asked to choose a side to join in the attacks. We believe this helps explain why the peak in the number of defacers does not coincide with the peak of defacements.

There was an exceptionally high volume of defacements targeting Russia on 25 May with 771 attacks caused by five defacers. However, 764 of these were claimed by a single attacker, who had compromised a server hosting 760 websites – an outlier that appears to be unique in our dataset (that outlier is removed from the graph for better visualisation). The peak of defacements of Ukrainian sites on 1 February (187 attacks by four defacers) does not appear to have a single cause and it did not lead to a sharp increase in defacers in the following days; we are not aware of any particular reason why attacks occurred on that day, several weeks before the invasion.

DDoS attacks lagged defacement by about a week. The number of both DDoS attacks and victims targeting Russia first increased on 2 March (6 days after the invasion) with 851 victims, 511 of them at around 6PM. The attacks peaked 4 days after with 1 137 victims on 6 March. High levels of activity continued through 23 March, with the biggest wave occurring at around 2PM on 8 March with 755 victims. Smaller waves continued regularly during the next few weeks. Regarding DDoS attacks hitting Ukraine, significant waves started around a week after Russia’s first big wave (some small spikes targeting Ukraine before Russia were insignificant and short-lived) with the first notable spike on 10 March with more than 526 victims. Attacks became prevalent during two weeks from 18 to 31 March: big waves were on 18 March at around 12PM, 1PM and 4PM with 257, 476, and 700 victims, respectively. Other big

Figure 2: Number of defacement and DDoS attacks hitting Russia and Ukraine over the period. The star marks the invasion day.

Figure 3: The number of web defacement and defacers (left); DDoS attacks and victims (right) targeting Russia and Ukraine by hour for four weeks around invasion day (indicated by red stars). The scales differ, but the larger the bubbles, the more attacks.
and medium waves also happened until the end of March, with the biggest peak occurred on 31 March when 1 296 victims were hit.

Overall, while significant numbers of defacements occurred immediately after the invasion, this rise was rather short-lived; the increase lasted for only around two weeks before returning to the previous stable state. We learnt from our interviews that defacers had run out of targets and had lost interest in targeting Russia or Ukraine (see more in §5 and §6). DDoS waves happened later, but occur in higher volumes and continued for longer. Yet, the increased volume only continued for about a month before declining sharply.

4.2 Attacks in the Global Context

Thus far we have discussed the number of attacks against Russia and Ukraine with significant changes connected to the war. However, these numbers are rather trivial when set against attacks occurring elsewhere in the world (Figure 4). Among 274 963 defacements being analysed, only 5 899 (2.15%) targeted the two countries (4 340 and 1 559 for Russia and Ukraine, respectively). We exclude sites hosted in the United States (US) by global scale vendors (Cloudflare, Amazon, etc., see Appendix C) but it is still the case that websites hosted in the US have consistently suffered the majority of defacements. Since the beginning of 2022, the US accounts for 26.95% defacements in total, followed by India (11.47%) and Indonesia (8.41%), while Russia and Ukraine only account for 1.58% and 0.57%, respectively. The top 10 countries account for 69.85% of all defacements; the same type of concentration is seen for DDoS, where the top 10 countries account for 70.49% of all victims. The US still dominates, with 24.68%, followed by Brazil (11.99%) and Bangladesh (8.10%). Ukraine accounts for 1.57%, while Russia lies eighth at 3.61%.

Our DDoS and defacements datasets are collected independently and do not affect each other, but they do show some correlation. Three among top 10 countries for defacements are also among top 10 for DDoS targets, namely the US, Germany, and Brazil. During the last two weeks of March 2022, the number of attacks (both defacements and DDoSs) rose at the global scale, with many defacements targeting the US and unprecedented volumes of DDoS attacks targeting Bangladesh. They both increased significantly on 21, 24 and 27 March, but dropped sharply on 22, 26 and at the end of March. It would be necessary to control for factors such as overall Internet activity before considering these correlations to be significant. The unusual peaks against Brazil in late June, for both defacement and DDoS attacks, are notable. Although Brazil is often ranked among the top 15 cybercrime hubs worldwide [57], we are unable to provide any convincing explanations for these.

A similar peak to that observed in Russia and Ukraine for DDoS attacks can be observed at the global scale in the aftermath of the invasion. We observe notable differences in defacements between the top-country global scale and the Russia-Ukraine scale. There was a short-lived decline in defacement attacks worldwide on invasion day (from around 1 400 to 1 000), while it peaked for Russia and Ukraine globally in March 2022 but quickly dropped to the pre-war level. The unusual peaks against Brazil in late June, for both defacements and DDoSs attacks, are notable. Although Brazil is often ranked among the top 15 cybercrime hubs worldwide [57], we are unable to provide any convincing explanations for these.

4.3 Key Actors in the Defacement Community

Key actors play central roles in underground and hacking communities; a small number of actors are often involved in many activities [39, 92]. Reflective DDoS attacks lack attacker identifiers, so we are limited to investigating whether we can identify 'key defacers'. Figure 5 shows the number of key active defacers and the
proportion of defacements they contributed. The defacement community is highly centralised: over 6 months, 10 defacers accounted for 30.06% of attacks, while the most active of them contributed 8.25% (around 22.7k). If we ignore the period before the invasion, conflict-related defacements show higher concentrations around a relatively small subset of actors: the top 10 targeting Ukraine accounted for 39.82%, while the most active was responsible for 9.10%. The effect is even more pronounced for defacements targeting Russia, where the numbers are 53.95% and 23.01%, respectively.

Among the most active defacers in the entire 6 months, two actively attacked both Russia and Ukraine when the war began: the 5th ranked 3rd & 4th, and the 9th ranked 4th & 1st, for attacking Russia and Ukraine, respectively. Some picked sides: the 8th ranked 7th for attacking Russia, while the 4th ranked 5th for targeting Ukraine. The 6th did not target either country at all. We found some ‘new faces’ e.g., the second most active defacer targeting Russia after the war began first appeared in mid-February, peaked on the invasion day, stayed significant for 3 days then declined quickly. Some ‘old faces’ performed many attacks against other countries but not Russia and Ukraine, yet they suddenly did that after the invasion, suggesting their choice of targets has been influenced.

These figures show that the conflict has caught the attention of some old players, but also drew in new ones. While some minor players at the global scale made a big contribution to the rise in attacks on Russia and Ukraine, the three most active defacers globally made only a trivial number of attacks against either country (less than 10). Our interviewees also suggested that cybercrime gangs post defacement trophies as a form of advertising while more established leaders prefer anonymity and are mainly financially motivated – ‘now they are busy with making money’, one explained.

4.4 Motives for Defacement Attacks

We do not seek to verify earlier findings about defacement motives in general (as discussed in §2), but to gain insights within the context of the conflict. Hence, we only look at the 5,899 attacks in our dataset targeting Russia and Ukraine (4,340 for Russia and 1,559 for Ukraine). We find 1,341 identical messages, then identify motives of each based on the message’s meaning. We consider there to be a political sentiment present and mark it as either supporting Russia or Ukraine if a support/objection is expressed e.g., ‘We stand with Ukraine’. We consider a message to be for amusement or gaining reputation if it consists of defacers’ signature e.g., ‘Hacked by Hero’ without a clear motivation, or just greetings to other peers. We exclude 1,278 reports (21.66%) where there is no evidence of the motive (e.g., empty or random messages). Messages advertising tools and hacking services or asking for ransom are considered financially motivated e.g., ‘contact me for tools and shells’, some relate to patriotic, sectarian or nationalistic conflicts. We label messages expressing favourite mottos or documenting the defacers’ mood as self-expression e.g., ‘Not much I want, hope my life will be better’.

We find diverse motivations, but although targeting Russia and Ukraine, most messages did not refer to the war at all. 2,723 (46.16%) were simply for amusement and reputation, 1,219 (20.66%) self-expression, while 143 (2.42%) related to other conflicts (such as Israel-Palestine), 58 (0.98%) related to patriotism and 89 (1.51%) were financially motivated (mainly from the two most active defacers at the global scale mentioned above, who did not change their targeting because of the war). However, some defacers did leave messages relating to the war: 286 (4.85%) supporting Ukraine, roughly 2.8 times higher than those supporting Russia at 103 (1.75%). Notably, some defacers support Russia, yet also defaced Russian sites saying that they wished to alert and help secure the systems (22 attacks) – ‘I have secured this domain, I love Russia’, was a message the third most active pro-Russia defacer left on a Russian website. Likewise, other defacers supported Ukraine yet defaced Ukrainian sites (12 attacks) e.g., ‘Hello Volodymyr Zelensky, I’m sorry to hack your site. I just wanted to tell you that people need a president like you. We support Ukraine’. Such messages could be signatures left over from exploiting tools, yet both our quantitative evidence and qualitative interviews suggest war-related messages during that period were explicitly intentional, as Russia or Ukraine were not targeted often before (see §4.1 and §6.3).

5 THE IT ARMY OF UKRAINE ACTIVITIES

The IT Army of Ukraine, which was stood up in an ad-hoc manner [82], has been promoting Russian targets most mornings with URLs and IP addresses posted on their Telegram channel. We analysed these announcements to explore the choice of targets, and the crossover with attacks seen in our datasets. We used regular expressions to extract IP addresses and domains, then combined data for the same domain (e.g., www.example.ru, smtp.example.ru). We found that the targets do not just have Russian and Belarusian ‘country codes’ (.ru, .su_, .by), but also generic ‘top level domains’ (e.g., .tv,.com). We excluded URL shorteners such as t.m,e.g, goo.gl and popular services such as youtube.com. In total, we found 3,845 targets promoted, including 2,921 IP addresses (59.58%) and 1,554 domains (40.42%), which are not necessarily distinct as targets may be described in both ways to make it easier to launch attacks: some DDoS tools require IP addresses be provided for target identification, and connecting to websites usually require hostnames.

5.1 Promoted Targets and Community Reaction

We recorded a large number of announcements and targeted domains in the first couple of weeks after the invasion. Posts began on 26th February, peaking on 27th February with 40 announcements and 45 domains promoted (IP addresses were not regularly included until later), see Figure 6. However, the volume of announcements quickly declined to consistently less than 10 per day after two weeks with some days (e.g., 24th or 26th April) having no announcement. The number of participants in the channel also dropped gradually to roughly 200k in early 2023, while it was previously over 300k.
While the number of announcements decreased, the number of promoted targets has steadily increased, particularly in May and June 2022, with announcements containing multiple targets. Activities at that time were unstable with targets being promoted less frequently; occasional periods of several days had no promoted targets at all. During the first couple of weeks, most targets were fresh (both domains and IP addresses), but a considerable proportion were re-advertised on multiple days. For example, during 4–6 May, all advertised IP addresses and most domains were re-posted. Along with frequent zero-target days, this suggests the organiser might run out of new targets or get bored with finding fresh ones.

Community reactions and engagement tell much the same story as with DDoS and defacement attacks (see Figure 7). While more targets were promoted in the last two months of the period, volunteers appeared to have shown intense activity in the first few weeks but to have largely lost interest thereafter. The decline in reaction was consistent across all types of engagement we could measure: views, emojis, forwards, and replies. Older announcements have had more time to accrue views as people scroll up the channel, but the emojis, forwards, and replies require user intent. We deduce that the figures reflect a genuine decline in engagement over time.

Besides targets promoted daily, the group also provided instructions such as tools and guidance to carry out DDoS attacks. We further looked at user engagement with these instructions when they were introduced into the channel. The group first provided material to hit Russian payment system on 9 March (two weeks after the invasion), attracting lots of engagement: 240k views, 2.6k emojis, 1.2k forwards, and 421 replies from 197 users. The next was on 1 April: while the number of replies and forwards was quite similar to the first, other kinds roughly halved. From mid-May to late June, instructions were posted more frequently (4 times), yet users were much less engaged (around 4 times less than the first in March), suggesting a loss of interest despite the operator’s efforts.

### 5.2 Crossover with Observed Attacks

The IT Army of Ukraine maintains a dashboard showing the status of targets, claiming many are down due to member actions. To determine whether the attacks involved reflected DDoS or whether web defacers were claiming successful attacks, we correlate the Telegram channel’s list of targets with our other datasets. Figure 8 shows the overlaps of promoted targets with defacements and DDoS attacks from 26 February 2022 (when the Telegram group became active). We consider an overlap with a defacement when either its URL or IP address matches one of the promoted targets, while for DDoS attacks, only the IP address is taken into account.

We observe very little overlap with defacement. Among 3845 promoted targets, there are only 59 valid matches (1.53%), including 7 domain matches (0.18%) and 52 IP matches (1.35%). Notably, no overlaps occur on the day targets are promoted, suggesting that defacers chose their targets themselves independently; these targets are largely unimportant and irrelevant to the conflict.

For DDoS attacks, among 291 promoted IP addresses, we observe 707 (30.86%) total overlaps, which is considerable. Unlike defacements, some are executed the same day they were promoted, yet these same-day overlaps still account for a trivial proportion of the global DDoS attacks targeting Russia observed in Figure 2, suggesting a little contribution from the channel to the overall attack landscape. We find many same-day overlaps in late March, early April and during May, peaking on 19 March 2022 with 22 victims overlapping. However, the crossover dropped quickly and became less frequent from late May, while many new targets were still advertised (as in Figure 6). This suggests a loss of interest by volunteers in executing attacks against the promoted targets.

---

YW_2022-06-06

Yandex, Government Data, Moscow Region Administration, Federal Service for Supervision of Transport, Gazeta.Ru News, Belgorod State University, and MultiChange.
5.3 The Choice of Targets

The targets announced on the IT Army channel are often associated with particular themes. These mini-campaigns are sometimes patterned around particular days of the week (e.g., food delivery services, entertainment, online news and propaganda are often targeted at weekends to maximise impact as people spend time at home online). Sometimes, the theme is set with re-advertised old targets of previous days, leading to wide variations in the number of new targets, particularly from May onwards (zero on some days), see Figure 6. Although subscribers can also contribute to the target lists, the group owner posts most of them; but during the last two months, they often stuck to promoting old targets via links to previous posts, which could be as simple as ‘we continue to work with yesterday's targets’. This could suggest that the group organisers had some difficulty in coming up with new targets.

We use categories linked with targets by default, but when they are unavailable, we identify them by considering their root domains (e.g., .tv and .gov likely belong to news and government websites). If they are generic (e.g., .ru, .com), we visit them (relaying via Russian IP addresses) or consult Internet archives if they are down. We found some targets were indeed down while they were previously active, suggesting that attacks might have been successful. For example, ksrf.ru (the constitutional court of the Russian Federation) has been down for some time, while data.gov.ru was both defaced and DDoSed. Although we can check targets’ online history on Internet archives, these may not be complete and so we did not attempt to measure what proportion had been taken offline for a substantial period. By examining the announcements, we were able to not only classify their categories, but also smaller sub-categories.

There is a wide range of categories, but five dominate 80.21% of all targets (see Figure 9). ‘News, media and propaganda’ (including TV broadcasting) has been consistently promoted since the war began, but only became the most common category in May when it overtook ‘IT solutions and services’. ‘Government and public services’, which includes military, state-owned websites, and public services for civilians such as parking and lighting (including governments imposed on occupied territories) has also been regularly targeted throughout, but they only grew rapidly towards the end of the period, making it the second most common category overall.

Ranked third is banking and financial systems, including banks, stock exchanges, electronic payment, accounting, credit checking services, trading, bidding, investment platforms and funding agencies. ‘IT solutions and services’ ranked fourth, including software solutions supporting governments, digital signature providers, and information security services such as DDoS Guard. This category was actively promoted early on but was targeted far less thereafter. Fifth is logistics, including airlines and aviation, travel, shipping, and food delivery. Other categories (by popularity) include online markets and stores (e.g., job markets, real estate, e-commerce, drug stores), manufacturers and trading (e.g., military footwear, wood and roofing materials), education, insurance, telecommunications (e.g., Internet providers), businesses and state companies (e.g., energy and steel manufacturers), discussion forums, entertainment (e.g., cinemas), and some non-governmental organisations.

6 DEFACEMENT COMMUNITY PERSPECTIVES

Our quantitative evidence reveals significant but fairly short-lived shifts of both defacement and DDoS attacks associated with the conflict, with distinct peaks on the invasion day. Yet, raw numbers produced by Internet measurement are notoriously difficult to interpret. Large variations and trends can be caused by any factor, which may interact with one another. As discussed above and has been previously identified in the literature on concentration, apparently large effects can be the result of limited individual actions [19]. To contextualise and support our quantitative observations, we extend our study to include qualitative research on what happened within the defacement community during wartime, and whether there was any impact on their choice of targets, motivations and behaviour patterns. As explained in §3, the interviews were conducted on a small scale and two participants recruited from two defacement teams cannot be generalised to the cybercrime ecosystem as a whole, yet they do give an instructive pair of case studies and provide some explanations as to why the peaks occurred.

6.1 Participant Demographics

Although formed around a strong regional identity, both participants indicated their teams had a wider support base from a range of other countries, with a small number of official members and a larger community of supporters. One team had 3 members, while the other team was larger, with 10 members. Members of both teams are quite young, ranging from 15 to 26, mostly high-school and college students. One team was all male, while the other was more diverse with one woman. One team knows each other in person (sometimes working at the same location), while the other team only meets online. Skill levels and histories of illicit behaviour varied considerably between members. Some claimed a long history of illegal activity, largely based around low-level cybercrime activities, while others had a much shorter history of involvement. The teams had a financial structure common to many online cybercrime groups – essentially acting as groups of friends involved in an entrepreneurial project selling cybercrime services.
6.2 Participant Cybercrime Activities

Both teams were involved in a range of activities associated with the low-level ‘cybercrime underground’ – a loose array of entrepreneurial schemes involving monetising pre-made attack scripts and providing illicit services – but defacement was the primary activity. They do hacktivism sometimes, mostly when their country was under attacks – ‘I did many attacks to Turkey and Iran when they are attacking Kurdish’, yet both have no political viewpoints on the Russia-Ukraine conflict. They also do hacking for profit – ‘I do not like white-hat at all’ – one works full-time at home, and does crypto mining on compromised systems; the other sells shells, databases and remote access to servers and networks – ‘In 1 month selling shells and database, I make $5 000; people pay very well for it’.

There is little evidence of a history of sustained targeted action for either group, with most defacements automated and untargeted – for amusement or competition. They started with manual and simple hacks on random targets, then automated scripts to scan potential victims based on registrars and ccTLD (e.g., ru, ua), then launched mass defacements with ready-made tools. Yet, these tools are themselves fairly unsophisticated – they can only deface sites with common unpatched vulnerabilities. Their attacks sometimes explicitly targeted public sector websites, largely when getting paid by others (or occasionally on a whim). Generally, as has been identified in previous research, attacking public infrastructure was seen as too high-risk in the context of running a lucrative service [20, 41].

Despite this, the groups did occasionally take part in overtly political campaigns, spreading anti-government messages, or targeting other countries when their own was perceived to be under attack.

We do not comment in detail on their technical skill levels – which are particularly irrelevant, given the nature of the contemporary industrialised cybercrime economy, but they are neither low-level cybercrime customers, nor highly-skilled pentesting experts or clandestine vulnerability researchers. Instead, they appear to be classic cybercrime entrepreneurs, whose own use of their tools outside a business context occasionally takes a political dimension. They do not fit the classic model of ‘hacktivists’ as they are rooted in the competitive defacement scene, but they demonstrate more political interest than many in the cybercrime ecosystem. Although neither group is based in Russia or Ukraine, they are precisely the sort of people, with access to cybercrime skills and tools, who have been targeted by calls for global participation in a cyberwar.

6.3 Participant Reactions to the Conflict

One participant claimed their friends had initially tried to support the IT Army of Ukraine but subsequently ‘moved to the Russian side due to not being treated well’. Opinions of the IT Army were generally low – while the targets were promoted by the group admins, the volunteers were apparently doing most attacks, yet ‘they were helping but they get no credits’. Neither team participated actively in the Telegram channel, but took part in self-directed mass defacement aligned with the war, largely for their own amusement. When the war began, although some teams within the defacement community did attest to political motivations (as noted above) the most prolific defacers did not engage by targeting either the countries but remained largely financially motivated (e.g., malicious SEO, selling shells) – ‘they are busy making money’, said a participant.

Both participants obtained war-related information from social media news, claimed to be influenced and changed targets specifically due to the conflict – ‘just the war, normally we do not target Ukraine’, one said. One team started defacing Russian sites with messages supporting Ukraine; they were at pains to argue that their issue was with the invasion, and that some of their Russian friends within the cybercrime community were similarly against the war. Sometimes they accidentally defaced news sites but they did not focus on any particular category (e.g., bank, government), which contextualises the relatively small overlaps between defacement and targets promoted by the IT Army of Ukraine. As for the other team, its owner was reported to be pro-Russian, but the participant executed all attacks for his team and described himself as neutral, attacking Ukraine for amusement – ‘they misunderstood telling that I am Russian side supporter, but that is not true’. The overall picture suggests that despite briefly being influenced by the war, neither team has a strong political viewpoint on the conflict, so their activities cannot properly be described as ‘hacktivism’ – ‘Defacers have no political reasons, they do it just for fun or as a hobby in the start, later on, they convert access to different targets on money’, one said.

Although involved in several kinds of cybercrime, when targeting Russia and Ukraine, both teams focused only on defacement; they saw it as an efficient way to spread messages and propaganda (also to advertise their services to make money). The participants mostly used traditional mass-defacement rather than specially-targeted methods. The use of such tools explains the immediate peaks and the lack of important sites being defaced. One team started defacing Ukrainian sites a couple of days after the invasion, while the other started targeting Russia within just a few hours after the war began and it was their sole focus for the next two days. While they were clearly influenced by the war, their interest was fleeting; it declined quickly as they ran out of targets and generally lost interest in the conflict – ‘I join fast and leave fast’. One participant became afraid of being arrested after their attacks were widely reported – ‘they stopped, then I stopped too’.

6.4 Wider Community Reactions to the Conflict

Our participants described their corner of the cybercrime underground as densely networked, with considerable communication even between competing teams. They share information, techniques and interact with others in a private Telegram channel. They reported co-operating on some campaigns (depending on the targets, sometimes they needed help), especially those linked with more political causes – ‘we trust each others because we are all criminals’, one explained. Traditionally, cybercrime communities have centred around underground forums, yet most of our participants’ communication had shifted to private channels such as Telegram and Signal, as also reported in the literature [46].

The defacement community was clearly influenced by the conflict; they frequently commented on the situation. There was a ‘call for hacking’ to encourage them to choose a side and participate in attacks – ‘Both Russian and Ukraine hackers contact me for help’, one said. While some decided to be pro-Russia or pro-Ukraine, some did not care and some decided to remain neutral – ‘they were asking me to choose a side, but I told them I am neutral and that I will not get involved in a war that is not mine’. Some pro-Russian defacers
actively targeted Ukraine, but also hit some Russian websites to help ‘secure’ them by demonstrating their vulnerabilities.

7 THE ROLE OF CIVILIAN HACKTIVISTS

Our findings suggest a significant yet fairly short-lived impact of the invasion on the activities of civilian and the so-called ‘volunteer hacktivists’ with links to low-level cybercrime activity. Government-backed attacks are hard to measure, at least for academics, but we are able to reliably measure non-governmental activity. Datasets covering the whole cybercrime underground are never available; while other attacks such as phishing and ransomware are not covered and the use of defacement and DDoS attacks cannot represent the entire ecosystem, we believe our measurements are instructive and the two independent datasets reflect the engagement of civilian and low-level criminals as they are common types of attack with the most mature economies – DoS-as-a-service, for example, can be purchased for as little as $5 per month [42]. Importantly, the triangulation of our quantitative evidence and qualitative interviews all point to the same story, a narrative that is missing from popular accounts of the conflict. We are not against the arguments made about the prevalence of state-sponsored attacks such as malware and phishing, but rather provide additional perspectives on the role of civilian and volunteer players in the continuing conversation.

Web defacement and DDoS attacks followed similar but differing trajectories at the start of the war. Our measurements indicate immediate effects on defacement activity, but they have very little overlap with the targets promoted by the IT Army of Ukraine. As high-profile websites (e.g., government) are difficult to compromise, the vast majority of attacks targeted harmless, defunct, or trivial sites with Russian domain names: victims were not intentionally targeted, and attacks were automatically executed by mass scanning with ready-made scripts. Through defacing websites, attackers express their supports to either Russia or Ukraine, but they do it mostly for fun; their attacks are often opportunistic rather than politically oriented. In many cases, web defacers did not stray far from their usual, financially motivated, cybercrime activity with adverts for their cybercrime services left on the defaced websites.

The rise in DDoS attacks did not happen until slightly later – a week after the invasion – and it is there that IT Army of Ukraine had considerable influence. At least 30% of targets promoted in the Telegram channel could be found as victims in our reflected DDoS attack database. However, our key result is that neither the increase in defacement nor DDoS attacks was long-lasting. Defacement activity was concentrated in the first couple of days; DDoS activity continued for longer (possibly due to the widespread availability of cheap DDoS-for-hire services), but was still fairly short-lived. Our qualitative data suggest that this can be partly explained by participants simply losing interest, although their targets were clearly influenced by the war. This is in line with other work suggesting that boredom is an important factor people leaving cybercrime [20].

Our findings fit with a more general pattern observed in the cybercrime ecosystem – one of a cybercrime underground increasingly characterised by an entrepreneurial, service-based economy which is becoming alienated from traditional hacker culture’s concerns with technical learning and dissent [3]. This is in line with a report on the motivations of cyberattacks in late 2022: hacktivism and cyber warfare only account for around 7%, while over 82% of attacks are cybercrime [70]. Rather than the committed, persistent attention needed for the prominent leak-based attacks observed throughout the war, the interest of low-level civilian ‘hacktivists’ from cybercrime groups appears to be fleeting, and easily diverted by trending news. This combination of deviant entrepreneurship and expressive juvenile delinquency suggests that the contemporary cybercrime underground is becoming far more like classic criminological portraits of offline street-level crime and disorder [67, 100]. Committed, effective hacktivism appears to be rather more separated from these low-level crimes than prevailing narratives might suggest.

It is true that some cybercrime-related activities are contributing to the war effort. Leaks, especially of high-profile datasets gathered from Russian public services, have consistently made headlines. Much as with ransomware, their low numbers and vast disparities in impact make them far less ‘measurable’. They may or may not be connected to ‘civilian engagement’, to ‘hacktivists’, to dissidents within the targeted organisations, to state actors, or to other groups. Equally, our argument is not that all commentary on the conflict is ill-founded, with more nuanced perspectives evident from some researchers, security professionals and analysts. We draw attention to Rid’s ‘Cyberwar Will Not Take Place’ [78], his later comments on Russia-Ukraine which echo our own [79], and security blogs such as those run by the Grugq [32], Paddy Kerley [50] and Microsoft [60].

8 CONCLUSION

The participation of civilian ‘hacktivists’ involved in cybercrime described in this paper – those which we believe are meaningfully measurable – are trivial acts of solidarity, teenage competition, and expressive delinquency. We have found little measurable evidence to suggest that they are making any real ‘hard’ contribution to the kinetic conflict, even in a major war between two nations with a long history of cyberwarfare. This analysis runs counter to widely-held popular narratives, which tend to collapse state hacking and political ‘hacktivism’ together with low-level cybercrime. This has broader consequences for our understanding of the role of civilian engagement and its capacities in nation-state conflict and for what we might expect to see in the future. It is particularly important to note that in cybercrime measurement data, what initially appears to be a large-scale movement can turn out on closer examination to be the result of a small number of actors changing their behaviour.

We are not dismissing these activities as completely unimportant or uninteresting. Rather, they engage in the domain of cultural expression, not hard power; they are part of an information war in a much older sense of the term. The history of the youth cultures which emerged during the cold war is instructive. For example, the punk movement alone was similarly engaged in graffiti and expressive protest, but had a real and lasting impact on global culture. These battles for ‘hearts and minds’ have long been considered to contribute to ‘soft power’; culture is seen by most nations as an important theatre of conflict and wider geopolitics. It is telling that much of the work of the IT Army of Ukraine has been the spreading of messages and propaganda. The people engaged with and encouraged to use off-the-shelf tools of the low-level cybercrime underground are employing those tools for much the same purposes, rather than participating on the ‘hard’ digital frontline.
ACKNOWLEDGMENTS

We are grateful to the 1877 Team and the Monday Team for their participation, and for providing constructive comments on the early manuscript. Their views and feedback have added valuable insights and perspectives into our findings and interpretations.

REFERENCES


C. Determining Attack Geolocation

Accurately mapping IP addresses to countries is challenging, as IP geolocation is not always stable and trustworthy [30]; providers prefer locating servers in countries with cheap hosting [96]. Attack geolocation might thus be determined differently by different archives e.g., ZONE-H may say an IP is in Germany, while ZONE-XSEC goes for Singapore and DEFACER-ID cannot tell. Geolocation services are more reliable at the country level [25], but this is only part of the truth as websites are nowadays commonly hosted on content delivery networks (CDNs). The original IP addresses are typically hidden and the geolocation is of the CDNs.

For example, a `.ru` website is supposed to be Russian, but it might be physically hosted in Vietnam, operated by a person living in Hong Kong, while proxied through Cloudflare with an IP address in the US. Relying on only one aspect might be risky, as both IP and domain can lie. We use data fusion to enhance the accuracy, prioritising: (1) top-level domain; (2) IP geolocation at collection time (MaxMind GeoIP2\(^7\) for web defacement, and a database we maintain based on Regional Internet Registry data for DDoS attacks; (3) geolocation of the AS hosting the IP address. If a website’s IP address belongs to a CDN, its geolocation is determined solely by ccTLD, as any geolocation of IP address or ASN will be unreliable.

The top three CDNs are Cloudflare, Amazon Web Services, and Akamai, serving around 89% of customers [43]. We ignore the trivial market shares of their competitors, but we count DDoS Guard as it is based in Russia, which may affect the infrastructure hosted there. We expect the four can cover nearly 90% of customers. In total, we found 4.87% of defacements are housed on these CDNs by 14262 prefixes as of the writing date: 1 698 of Cloudflare; 7 483 of Amazon Webservice; 5 056 of Akamai; and 25 of DDoS-Guard (these prefixes and AS number mappings are collected on Hurricane Electric Internet Services. For defacements, we prefer ccTLD over IP geolocation as attackers likely target websites in a country by massively scanning domain ccTLD (e.g., `.ru`, `.ua`) rather than checking if IP addresses are hosted in that country.

Accurate measurement of frequent ccTLDs used in Russia and Ukraine is complex; many Ukrainian firms use Russian services and vice versa. The most frequented domain used in Russia is reported `.ru` [85]. We cannot find a similar report for Ukraine, yet we believe incorporating ccTLDs with IP and AS geolocation is fairly reasonable as choosing targets based on ccTLDs is a straightforward way used by low-lever cybercrime actors.

D. Data Defacement Collection

Data completeness and reliability are critical for longitudinal measurements to measure real trends, not artefacts. Scraping complete snapshots, especially at ZONE-H, is non-trivial, and has not been guaranteed in prior work. Some have attempted to purchase ZONE-H snapshots [38], but this is not a sustainable solution, and is ethically questionable. Defacement archives are public, but gathering them at scale is still challenging as (1) ZONE-H adopts text CAPTCHA to prevent bots, (2) its dashboards set a limit of 50 pages with older data.

\(^7\) GeoIP2 is freely accessible at https://maxmind.com/. It offers both free and paid licenses, with the paid one being slightly more accurate and up-to-date. It claims to provide over 99.8% country-level and over 60% city-level accuracy, yet that varies from country to country e.g., 79% for Russia and 65% for Ukraine, within a 250km radius.
Algorithm 1 Semi-automatic defacement validation

1: procedure validate_defacements
2:   for each a ∈ verifiedAttack() do
3:     a.status ← 0
4:   end for
5:   for each a ∈ filteredAttacks() do
6:     a.status ← 1
7:   end for
8:   P ← pendingGroups()
9:   V ← verifiedGroups()
10: for each p ∈ P do
11:     T ← {}
12:     for each o ∈ V do
13:         d ← levenshtein(p, o)
14:         T ← topSimilar(d, T)
15:     end for
16:     showAssistance(T)
17:     s ← manualDecision()
18:     for each a ∈ P do
19:         a.status ← s
20:     end for
21:     if isValidated(s) then
22:         V ← V ∩ p
23:     end if
24: end for
25: end procedure

hidden, and (3) on-hold records may not appear on the homepage promptly, leading to potential misses when visiting the dashboard infrequently. The only way to get a complete scrape is by iterating through all submission IDs (this generated non-trivial workload) with the IDs of valid and invalid reports often mixed. Dealing with these issues, plus IP blacklisting and bot prevention mechanisms, is the main challenge to scraping. We responded by (1) developing an efficient text CAPTCHA solver for Zone-H utilising image-processing techniques, (2) routing our scraper through multiple proxies, and (3) carefully iterating through all submission IDs in turn. We stored raw data in a database to avoid unnecessary requests in the future.

E. Validating On-hold Defacements

The validation of reported defacement is not clearly documented. While Zone-H reports are kept on hold until manually verified by staff, Zone-Xsec, Defacer-Pro, and Haxor-ID use automatic validation, insisting that messages left on the defaced pages contain keywords linked to hacking activities (e.g., ‘Hacked by Me’). All archives specify validation status explicitly when visiting defacement by ID, except Zone-Xsec (where it only shows in the dashboard with a 50-page limit). Submissions to Zone-H manually reviewed by staff may be slow to verify, while automatic verification of the others is error-prone. Unverified records may be kept on hold forever, leading to incomplete longitudinal data. Consequently, collecting only defacements shown in the dashboard is inadequate, thus a complete defacement dataset is challenging to gather. To enhance data completeness and reliability, we perform a semi-automatic validation to confirm if unverified defacements are in fact valid.

Our strategy is shown in Algorithm 1. First, we consider reports verified by archives to be valid. Second, messages on the defaced pages of on-hold ones are used to decide their validity (as defacers often leave their signature to gain reputation e.g., ‘Hacked by CoolHacker’). Among those, we filter for submissions where the message specifically contains common hacking terms: ‘hacked by’, ‘h4ck3d by’, ‘h4ck3d by’, ‘pwn3d by’, ‘pwn3d by’, ‘pwn3d by’, ‘pwned by’, ‘pwned by’, ‘own3d by’, ‘owns3d by’, ‘touched by’, and ‘kissed by’ then consider to be valid if those messages also include defacers’ handles. For example, a message ‘This website was hacked, contact me t.me/coolhacker’ posted by a notifier ‘CoolHacker’ is considered valid. This method is looser than an exact comparison with ‘Hacked by CoolHacker’, but is still accurate; all 100 randomly checked examples were correct. Third, the remaining reports are manually validated by looking for the defacers’ signature. While some messages are obvious, some are complicated to decide. To speed up the annotation, candidates are grouped by handles and messages (normalised, redundant spaces removed); for each case, validated defacements with the most similar messages are suggested to the annotator. Levenshtein distance is used to estimate the similarity between two text s1 and s2, which is helpful as messages are often slightly modified from existing templates. If no message is found (instead images, iframes, or javascript), or it is challenging to spot leftover signatures (some are messy and hard to read), a web browser pops up to display the defaced page and assist annotators.

This assistance considerably reduces the required human effort. One difficulty with verification is dynamic redirects to a defacer’s webpage, as this can be modified later; when the defacer’s page is down, the defacement appears to point to a non-existent site, but a careful check could reveal evidence of a defacer’s domain name. We ignore cases that lack evidence, so as to ensure that those flagged ‘valid’ are indeed valid. We do not use complex machine learning techniques as message texts contain lots of noise; given a small number of samples (10K), ML does not seem to be more effective than a rule-based approach. Sometimes defacements appeared to be already verified at the time of collection, but became invalid afterwards; we re-validate all of them months after collection to make sure their status has been finalised by the archives.

F. Unifying Defacements and Defacers

Defacements are hashed based on (1) reporting date, (2) original handle of defacers, (3) root domain of victims, and (4) original message left on the defaced pages. The reporting date may be problematic if the defacer waits a few days to resubmit to other archives; but excluding it may lead to spotting matched records where they were submitted on different days due to repeat victimisation. As the unifying process needs to be precise and the number of unique defacers is just a few thousand, fully automatic analysis, for example using machine learning, is not appropriate. We therefore adopt a semi-automated approach combining automated handle similarity analysis with manual review of messages to identify types in handles. First, similar pairs of handle are extracted using the Levenshtein metric, constrained so that the distance does not exceed 25% of the handles’ length. Then, 10 sample messages left by defacers in each pair are automatically shown to the annotator; if the messages are semantically closed enough, the pair can be
unified under a single nickname. The decision relies on various aspects: message patterns, stylometry, synonyms, the inclusion of handles in the messages, typos in the handles, team, nationality, messages’ semantics, language, and even the rarity of handles (rare usernames like ‘cj2ks’ are more likely to be used by a single person, while common ones like ‘glory’ are more likely to be shared by multiple individuals [56]). Messages left are diverse: some are identical, some are relatively similar, some are rather distinct and contain the defacers’ name, and some come with phone numbers which may link to their country. We only confirm where there is sufficient evidence, and leave uncertain pairs unmatched. Many different typos occur e.g., missing characters, order of characters, case-sensitive. We find many different handles which leave similar messages used across different archives, making it hard to identify overlaps without unifying handles. When we are unsure, we may ask our participants specifically.

G. Interview Material

Before the interviews, we asked participants to review the information sheet and agree to our terms that their responses will be kept for some time for transcription, but will be then permanently deleted. Any of their inquiries about us and the interview process will be addressed before proceeding to the main questions below.

Demographic and Organisation.

(1) Do you deface websites as an individual or as a team? If so, what is your role? Do you know each other offline? How was the team formed? When did you join the team? What does your team name stand for?

(2) What is your gender, employment status, level of education, and age range (also of your team members)?

(3) What geographic area of the world are you located in e.g. continent or country? A specific location is not required.

Cybercrime Activities.

(1) How long have you been doing hacking? What type of hacking activities do you or your team often do? And how long have you been doing defacement in particular?

(2) Have you defaced websites and posted these on online archives? If yes, which archives do you often use, why do you choose them? Do you submit to multiple archives?

(3) Do you have a hacking community for discussion? Which forum or other platforms do you mainly use for that?

(4) Why do you deface sites? What do you gain from defacing websites (motivation and benefit)?

(5) What are the popular online archives you use to report defacements? Do you mirror defacements on multiple archives?

(6) What types of target do you deface? How has this changed over time? Do you particularly target important sites like government, military, or education? And why?

(7) Are there any targets that you would avoid? And why?

(8) Do you automate any part of website defacement?

(9) Do you know other defacers? Do you exchange contact? Are they your friends? Do they do other types of crime?

(10) How did you start defacing websites? How old were you when you did (age range fine)? How did you learn the skills? How have your methods changed over time?

(11) What vulnerabilities do you often exploit? What types of websites often have those (e.g., Wordpress, Drupal)?

(12) Do you get involved in other types of cybercrime such as DDoS, or only web defacement?

(13) How do you decide what to post on defaced sites e.g., team introduction, politics?

(14) Do you cooperate with other hackers? Have you faced any challenging target that you could not get through where you needed others’ help?

(15) How morally wrong/serious do you think defacement is? How serious do you think the police consider defacement to be? Do you have any fear of being targeted by police?

Perspective on the Conflict.

(1) Do you have a particular political point of view? What is your opinion about the conflict? Do you support a particular side, or you are neutral?

(2) Where do you source your news/information about the conflict? How has the conflict impacted your activities, and the activities of other hackers in your community?

(3) Do you see a cyberwar? Were you and your community aware of the conflict beforehand or of any preparations?

(4) Do you know the IT Army of Ukraine? Are you a member of that groups? If yes, why did you participate?

(5) Have you committed other types of cybercrime linked to the conflict, or just web defacement?

(a) If they targeted either Russia or Ukraine:

(i) Did you change your target due to the war, or for other reasons? When do you change it e.g., right after the war began? And were you affected by attacks waves against Russia and Ukraine?

(ii) Do you shift your activities to target only Russia or Ukraine websites, or are you still doing other types of hacking as usual?

(iii) You only targeted Russian or Ukraine websites for a few days, then it appeared that you lost interest. Is that correct, and why?

(b) If they did not target either Russia or Ukraine:

(i) Why did not you target Russia or Ukraine?

(ii) What was the main type of hacking activities that you focus on during that time?

(6) Do you know where these teams come from (send out a list of top teams we would like to know about)?

(7) We tried to contact other teams but they do not respond. Can you introduce us to them? Do you know why they do not want to be interviewed? The most prolific actors did not target either Russia or Ukraine; do you know why?

Future Research Collaborations.

(1) Would you like your team to appear in our publication? You can remain anonymous if you wish.

(2) Are you interested in future collaboration with us? Can we keep in touch and contact you if we have further questions?

(3) Would you like to read through our manuscript and give us comments to avoid misunderstandings and to ensure the paper does not leak any unwanted information?