AUGUST 2022

DETAILED REPORT

Quarterly Adversarial Threat Report

Ben Nimmo, Global Threat Intelligence Lead
David Agranovich, Director, Threat Disruption
Margarita Franklin, Director of Public Affairs, Security
Mike Dvilyanski, Head of Cyber Espionage Investigations
Nathaniel Gleicher, Head of Security Policy

TABLE OF CONTENTS

PURPOSE OF THIS REPORT

Our public threat reporting began about five years ago when we first shared our findings about coordinated inauthentic behavior (CIB) by a Russian influence operation. Since then, we have expanded our ability to respond to a wider range of adversarial behaviors as global threats have continued to evolve. To provide a more comprehensive view into the risks we tackle, we've also expanded our regular threat reports to include cyber espionage, inauthentic behavior, and other emerging harms — all in one place, as part of the quarterly reporting series we're testing. In addition to sharing our analysis and threat research, we're also publishing threat indicators to contribute to the efforts by the security community to detect and counter malicious activity elsewhere on the internet (See Appendix).

We expect the make-up of this report to continue to evolve in response to the changes we see in the threat environment and as we expand to cover new areas of our Trust & Safety work. This report is not meant to reflect the entirety of our security enforcements, but to share notable trends and investigations to help inform our community's understanding of the evolving security threats we see. During some quarters, our reporting may focus more on a particular adversarial trend or tactics we see emerge across different threat actors. During other quarters, we may dive into an especially complex investigation or walk through a novel policy application and related threat disruptions. We welcome ideas from our peers across the defender community to help make these reports more informative, and we'll adjust as we learn from feedback.

For a quantitative view into our Community Standards' enforcement, including content-based actions we've taken at scale and our broader integrity work, please visit Meta's Transparency Center here: https://transparency.fb.com/data/.

KEY FINDINGS

- Our quarterly threat report provides a comprehensive view into the risks we see across multiple policy violations including Coordinated Inauthentic Behavior (CIB), cyber espionage, inauthentic behavior and other emerging harms, like mass reporting.
- We took action against two cyber espionage operations in South Asia. One was linked to a group of hackers known in the security industry as Bitter APT, and the other, APT36, to the state-linked actors in Pakistan. More here.
- As part of disrupting new and emerging threats, we removed a brigading network in India, a mass reporting network in Indonesia and coordinated violating networks in Greece, India and South Africa. More here.
- Under our Inauthentic Behavior policy against artificially inflating distribution, we took down tens of thousands of accounts, Pages and Groups around the world. Our manual investigations around the Philippines election allowed us to build automated enforcement systems to defend against this sort of activity globally and at scale. More here.
- We also removed three networks engaged in CIB operations, including one network linked to a PR firm in Israel and two separate troll farms – one in Malaysia targeting domestic audiences and one in Russia targeting global discourse about the war in Ukraine. We included an in-depth threat research and analysis into the Russian network linked to the self-proclaimed entity CyberFront Z and individuals associated with past activity by the Internet Research Agency (IRA). More here.

01

Removing two cyber espionage networks from South Asia

Cyber espionage actors typically target people across the internet to collect intelligence, manipulate them into revealing information and compromise their devices and accounts.

As part of these latest disruptions against both networks, we took down accounts, blocked their domain infrastructure from being shared on our services, and notified people who we believe were targeted by these malicious groups. We also shared information with security researchers and our industry peers so they too can take action to stop this activity. We have included threat indicators, including malware hashes and command and control infrastructure, in the Appendix to this report, to enable further research and detection by the security community.

Bitter APT

We took action against a group of hackers — known in the security industry as Bitter APT — that operated out of South Asia, and targeted people in New Zealand, India, Pakistan and the United Kingdom.

While this group's activity was relatively low in sophistication and operational security, it was persistent and well-resourced. Bitter used various malicious tactics to target people online with social engineering and infect their devices with malware. They used a mix of link-shortening services, malicious domains, compromised websites and third-party hosting providers to distribute their malware. Our platform was one of the elements of the broader cross-platform cyber espionage operation.

We identified the following *new and noteworthy* tactics, techniques, and procedures (TTPs) used by this threat actor across the internet:

- **Social engineering:** Bitter relied on fictitious personas, posing as attractive young women, journalists or activists, across the internet to build trust with the people it targeted to trick them into clicking on malicious links or downloading malware. Rather than indiscriminately targeting people with phishing, this group typically invested time and effort in establishing connections with its targets through various channels, including email.

- **iOS application:** Our most recent investigation found Bitter deploying a chat application for iOS that users could download via Apple's Testflight service for developers to help them beta-test their new applications. This meant that hackers didn't need to rely on exploits to deliver custom malware to targets and could utilize official Apple services to distribute the app in an effort to make it appear more legitimate, as long as they convinced people to download Apple Testflight and tricked them into installing their chat application. We don't have any visibility into whether this app contained malicious code and assess that it may have been used for further social engineering on an attacker-controlled chat medium. We reported our findings to Apple.

- **Android malware:** We found Bitter using a new custom Android malware family we named Dracarys. Notably, it used accessibility services, a feature in the Android operating system to assist users with disabilities, to automatically click through and grant the app certain permissions without the user having to do it. Bitter injected Dracarys into trojanized (non-official) versions of YouTube, Signal, Telegram, WhatsApp, and custom chat applications capable of accessing call logs, contacts, files, text messages, geolocation, device information, taking photos, enabling microphone, and installing apps. While the malware functionality is fairly standard, as of this writing, malware and its supporting infrastructure has not been detected by existing public anti-virus systems. It shows that Bitter has managed to reimplement common malicious functionality in a way that went undetected by the security community for some time.

- **Adversarial adaptation:** This group has aggressively responded to our detection and blocking of its activity and domain infrastructure. For example, Bitter would attempt to post broken links or images of malicious links so that people would have to type them into their browser rather than click on them — all in an attempt to unsuccessfully evade enforcement.

APT36

We took action against a group of hackers in Pakistan — known in the security industry as APT36 — that targeted people in Afghanistan, India, Pakistan, UAE, and Saudi Arabia, including military personnel, government officials, employees of human rights and other non-profit organizations and students. Our investigation connected this activity to state-linked actors in Pakistan.

While this group's activity was relatively low in sophistication, it was persistent and targeted many services across the internet – from email providers to file-hosting services to social media. APT36 used various malicious tactics to target people online with social engineering to infect their devices with malware. They used a mix of malicious and camouflaged links, and fake apps to distribute their malware targeting Android and Windows-run devices.

We identified the following *new and noteworthy* tactics, techniques, and procedures (TTPs) used by this threat actor across the internet:

- **Social engineering:** APT36 used fictitious personas — posing as recruiters for both legitimate and fake companies, military personnel or attractive young women looking to make a romantic connection — in an attempt to build trust with the people they targeted.
- **Fake, spoofed and real websites:** APT36 deployed a wide range of tactics, including the use of custom infrastructure to deliver their malware. Some of these domains masqueraded as photo-sharing websites or generic app stores, while others spoofed the domains of real companies like the Google Play Store, Microsoft's OneDrive and Google Drive. Additionally, this group used common file-sharing services like WeTransfer to host malware for short periods of time.
- **Camouflaged links:** This group used link-shortening services to disguise malicious URLs. They also used social cards and preview sites — the online tools used in marketing to customize what image is displayed when a particular URL is shared on social media — to mask redirection and ownership of domains APT36 controlled.
- **Android malware:** APT36 didn't directly share malware on our platforms, but rather used the above tactics to share malicious links to sites they controlled and where they hosted malware. In several cases, this group used a modified version of commodity Android malware known as XploitSPY available on Github. While XploitSPY appears to have been originally developed by a group of self-reported ethical hackers in India, APT36 made modifications to it to produce a new malware variant we call LazaSpy. These changes included adding technical checks to only run malicious functionality if the target's device

was located in either Pakistan or India and wasn't an emulator — likely in an attempt to avoid scrutiny by security researchers. Our malware analysis of several samples, however, found that APT36 failed to properly implement this functionality. This threat actor is a good example of a global trend we've seen where low-sophistication groups choose to rely on openly available malicious tools, rather than invest in developing or buying sophisticated offensive capabilities. As such, APT36 is known for using a range of different malware families, and we found that in this recent operation it had also trojanized (non-official) versions of WhatsApp, WeChat and YouTube with another commodity malware family known as Mobzsar or CapraSpy. Both malware families are capable of accessing call logs, contacts, files, text messages, geolocation, device information, photos and enabling microphone.

02

Removing “emerging harms” networks[1]

While our threat disruption work began with tackling inauthentic operations where people hide who’s behind them, we have also seen authentic actors engage in adversarial and harmful behaviors on our platform and across the internet. This section of the report details how our thinking about this adversarial space has evolved and what steps we’re taking to stay ahead.

This work began following the US 2020 elections. Since then, we have been working with teams across Meta to expand our network disruption efforts to new areas so we can address threats that come from groups of authentic accounts coordinating on our platform to violate our Community Standards and cause harm.

Here is how our efforts against emerging harms work in a nutshell: Our cross-functional security teams who work on these adversarial behaviors act as a “threat intelligence incubator”: they identify and study specific adversarial behaviors and then develop tailored policies and enforcement protocols to take action against those behaviors. Next, they investigate and disrupt networks, carefully scoping the enforcements to ensure we avoid over-enforcing and silencing innocent users. Over time, as we learn more and understand the nature of the threats more clearly, our goal is to transition each of these individual targeted efforts from the disruption-only phase to also include scaled automated detection. We do this by feeding the common tactics and techniques we see these networks rely on into our scaled detection and enforcement systems. To inform the community about our ongoing efforts, we’ve begun publicly reporting and sharing our findings with security researchers and our industry peers.

Here are the new problem areas we’ve been working to tackle:

- **Mass reporting**, in which groups of people coordinate to abuse our reporting tools by falsely reporting people in an attempt to silence them;

[1] *The threat disruption program focused on emerging harms outlined in this section was developed and launched by a multidisciplinary team working across Meta, including Artemis Seaford and Alberto Fittarelli who led this effort.*

- **Brigading**, in which groups of people coordinate to harass people on our platforms in attempt to intimidate and silence them;
- **Coordinated Violating Networks**, in which groups of people work together to break the rules outlined in our Community Standards.

These individual efforts are currently in various stages of maturity and will continue to move from disruption-only enforcements to adding automated detection as part of our disruption toolbox for each of the new problem areas we tackle. We share more details on our progress and specific enforcement examples in the following subsections.

Mass reporting

What is it? *Under our Inauthentic Behavior policies, we remove activity when we find adversarial networks that coordinate to abuse our reporting systems to get accounts or content incorrectly taken down from our platform, typically with the intention of silencing others. We began developing and implementing this policy in early 2021.*

As an example, in Q2 of 2022, we removed a network of about 2,800 accounts, Groups and Pages in Indonesia that worked together to falsely report people for various violations, including hate speech, impersonation, terrorism and bullying, in an attempt to have them and their posts wrongfully removed from Facebook. Most of these reports focused on people in Indonesia, primarily within the Wahhabi Muslim community. To conceal their activity and avoid detection, the individuals in this network would replace letters with numbers when posting about their targets. They, at times, created fake accounts that impersonated real people and then used them to report authentic users for impersonation.

What factors we consider when investigating mass reporting:

- Coordination signals
- High volume of reports
- Misleading & abusive nature of reports (e.g. reporting innocuous posts as violating)

NOTE: *We expect malicious groups will keep trying to break our rules and evade our detection. We continue to evolve our defenses in response to adversarial adaptation we see. To avoid tipping off these groups, we will not be sharing the exact thresholds and precise signals we rely on to tackle this abuse.*

Brigading

What is it? *Under our Bullying and Harassment policies, we remove activity when we find adversarial networks that work together to engage in repetitive behavior, often in the form of sending direct messages to their targets, or mass-commenting on their posts. The behavior is usually intended to overwhelm, harass or silence the target. We began developing and enforcing this policy in 2021.*

For example, in Q2 of 2022, we took down a brigading network of about 300 accounts on Facebook and Instagram in India that worked together to mass-harass people, including activists, comedians, actors and other influencers. This network was active across the internet, including Facebook, Instagram, YouTube, Twitter and Telegram. On our apps, the individuals behind this activity relied on a combination of authentic and duplicate accounts — many of which were disabled for violating our rules against hate speech and harassment by our scaled, automated systems. These accounts would call on others to harass people who posted content that this group deemed offensive to Hindus. The members of this network would then post high volumes of negative comments under the targets' posts. In response, some people would hide or delete their posts leading to celebratory comments claiming a "successful raid."

What factors we consider when investigating brigading:

- Repetitive targeting to harass or silence people, usually with unsolicited messages or comments
- Coordination signals
- High volume of activity
- Efforts to evade enforcement

NOTE: *We expect malicious groups will keep trying to break our rules and evade our detection. We continue to evolve our defenses in response to adversarial adaptation we see. To avoid tipping off these groups, we will not be sharing the exact thresholds and precise signals we rely on to tackle this abuse.*

Coordinated Violating Networks

What is it? *Under our Account Integrity policies, we remove coordinated violating networks when we find people — whether they use authentic or fake accounts — working together to violate or evade our Community Standards. We began developing and enforcing this latest policy in 2021.*

Our Community Standards govern what is and isn't allowed on our services. We have automated detection and manual review systems to address content violations, and we publicly report the quantitative results of this work at global scale on a quarterly basis.

However, in some cases, we've observed tightly organized groups working together to violate our rules while taking steps to evade enforcement — hence combining adversarial behavior with content-level violations of our Community Standards, like incitement to violence, hate speech, bullying and harassment, or misinformation. In these cases, the potential for harm caused by the totality of the network's activity far exceeds the impact of each individual post or action.

In response to organized groups relying on authentic means to break our rules, we created the Coordinated Violating Networks (CVN) policy. It goes beyond our existing responses to content-level violations and enables us to take action against entire networks — whether these are centralized adversarial operations or more decentralized groups — as long as they work together to systematically violate our policies (see examples further in this subsection).

Since we've developed this latest enforcement lever, we've enforced against networks with widely varying aims and behaviors. It included: groups coordinating harassment against women, decentralized movements working together to call for violence against medical professionals and government officials, an anti-immigrant group inciting hate and harassment, and a cluster of activity focused primarily on coordinating the spread of misinformation.

As mentioned at the start of *Section 2*, we rely on our tailored enforcement protocols to take action against these coordinated violating networks by manually investigating networks and then carefully determining the scope of each network disruption to help us avoid over-enforcing and silencing innocent users. Over time, as we learn more and further understand the nature of the threats, we will begin working to add automated detection to our toolbox against this type of abuse.

What factors we consider when investigating coordinated violating networks:

- Coordination signals that might show an organized group of people directly working together under centralized directions, or individuals that are part of a decentralized online community working towards a shared goal through similar activities
- Systematic violations of our Community Standards
- Efforts to evade enforcement

NOTE: *We expect malicious groups will keep trying to break our rules and evade our detection. We continue to evolve our defenses in response to adversarial adaptation we see. To avoid tipping off these groups, we will not be sharing the exact thresholds and precise signals we rely on to tackle this abuse.*

Here are some examples of our latest CVN enforcements:

Greece: We removed two clusters of accounts and Pages on Facebook and Instagram that worked together to repeatedly violate our policies against misinformation, hate speech and incitement to violent overthrow of the government. They were associated with two conspiracy groups: the "Guardians of the Constitution" and the "Holy Declarationists" who position themselves as the true protections of the Constitutions and argue that the Greek government has no constitutional authority. The individuals behind this activity used authentic and duplicate accounts to manage Groups and Pages, some of which were enforced against by our review systems for various content violations, including incitement to violence. They targeted politicians, judges, doctors, journalists and educators with calls to violence and harassment. According to public reporting, individuals connected to this activity were linked to the kidnapping of a high school principal for enforcing COVID-19 checks. They brought him to the police to report him for breaching the constitution, which led to the arrest of the kidnappers.

- **Most common content violation types by this network:** Violence & incitement; hate speech; bullying and harassment; misinformation.

India: We removed several clusters totalling about 2,000 accounts, Pages and Groups on Facebook and Instagram that targeted women in India with sexualizing content and harassment. The people behind each cluster of activity used authentic and duplicate accounts to manage Pages and Groups and flock to female users' accounts with uninvited content, including nudity, sexual solicitation and hate speech. In at least one case, an account targeted at least 700 people.

- **Most common content violation types by this network:** Sexual solicitation; hate speech; bullying and harassment.

South Africa: We removed several clusters totalling about 200 Facebook accounts, Pages and Groups that coordinated the harassment of migrants from other countries in Africa. Some of them organized under the brand "Operation Dudula," which according to public reporting is used by some in South Africa to campaign against undocumented foreign workers and blame them for poverty. Some members of Operation Dudula have been publicly linked to street violence. On our services, this particular cluster of activity included Pages and Groups that called for attacks against migrants, organized offline marches and events, and celebrated reports of violence. Our automated systems detected and removed much of this content. Likely in response to our detection and in an attempt to evade enforcement, the people behind this activity used duplicate and inauthentic accounts to manage their Pages and Groups. While we aren't banning all Operation Dudula content, we're continuing to monitor the situation and will take action if we find additional violations to prevent abuse on our platform and protect people using our services.

- **Most common content violation types by this network:** Violence and incitement; bullying and harassment; hate speech; misinformation.

03

Removing Inauthentic behavior

What is Inauthentic Behavior? Inauthentic behavior (IB), as detailed in our Community Standards, is an effort to mislead people or Facebook about the popularity of content, the purpose of a community (i.e. Groups, Pages, Events) or the identity of the people behind it. It is primarily centered around amplifying and increasing the distribution of content, and is often (but not exclusively) financially motivated.

IB operators typically focus on quantity rather than the quality of engagement. For example, they may use large numbers of low-sophistication fake accounts to mass-post or like their content — be it commercial, social or political. They often use tactics similar to other large-scale online activities, like spam.

This behavior pattern distinguishes IB from Coordinated Inauthentic Behavior (CIB) where operators invest in mimicking human social activity as closely as possible. Think of an elaborate fictitious journalist persona or a Middle-East-focused think tank managing multiple online accounts and websites to support their cover story, while trying to build trust with their targets and promote particular narratives. IB operators, on the other hand, can sometimes involve the use of fake accounts, but we typically see little attempt to obfuscate their identity from Facebook and only the most superficial attempts to construct a false identity.

Both of these violations are serious issues and we enforce against both, but we rely on very different tools that respond to their distinct behaviors. CIB networks typically require manual, expert investigations to uncover deception, whereas the relatively non-complex and repetitive nature of IB makes it particularly vulnerable to scaled detection and automated enforcement systems.

This approach allows us to learn and improve our defenses in response to adversarial adaptation across both violations, while removing IB clusters of activity at scale, no matter whether they aim to promote celebrity gossip, political news or clickbait with an aim to amass an audience. Importantly, constantly refining our scaled enforcement and detection allows us to disrupt IB

clusters sooner, reducing their ability to build an audience. If CIB operations turn to IB tactics to massage their engagement figures, they may also be taken down by our automated systems without manual investigation — freeing our expert investigators to focus on emerging, sophisticated threats.

Our recent work in the Philippines ahead of its election is a good example of this approach to combining expert investigations and scaled detection and enforcement. It is a model we're hoping to extend to other areas of our threat disruption work as we continue building our understanding of the threat environment worldwide.

In focus: Philippines

Manual investigations and disruptions:
Ahead of the Philippines election, our investigative teams took down about 10,000 accounts for violating IB policy. They used IB tactics to inflate the distribution of content that included election-related posts, including some that used politics as a spam lure at the time when people were interested in following these topics. Through this threat intelligence work, we continued to work on identifying repetitive patterns of behavior that are most characteristic of IB clusters in the region.

Automated detection at scale:
Based on these and earlier insights, we were able to automate the detection of these IB patterns to complement manual investigations. As a result, we identified hundreds of IB clusters in the Philippines and took action against over 15,000 thousand accounts after expert review to ensure we don't over-enforce. On average, these spammy clusters were less than six months old when we disabled them.

Automated enforcement:
In addition to manual disruptions and automated detection, we also focused on automating enforcement against some of these IB patterns, based on the most reliable signals derived as part of our election preparation work in the Philippines. As a result, our disruption systems were able to tackle specific types of high-confidence, repetitive inauthentic behavior in the Philippines and globally. Specifically, we took down more than 50,000 accounts engaged in IB, with about 10% originating in or targeting the Philippines and the rest coming from at least 100 different countries. Most of them were less than two months old when we caught them.

04

Coordinated inauthentic behavior (CIB)

We view CIB as coordinated efforts to manipulate public debate for a strategic goal, in which fake accounts are central to the operation. In each case, people coordinate with one another and use fake accounts to mislead others about who they are and what they are doing. When we investigate and remove these operations, we focus on behavior rather than content — no matter who's behind them, what they post or whether they're foreign or domestic.

Continuous CIB enforcement: We monitor for efforts to come back by the networks we previously removed. Using both automated and manual detection, we continuously remove accounts and Pages connected to networks we took down in the past.

Malaysia

We removed 596 Facebook accounts, 180 Pages, 11 Groups and 72 Instagram accounts for violating our policy against coordinated inauthentic behavior. This network originated in Malaysia and targeted domestic audiences in that country.

The individuals behind it ran a troll farm — a coordinated effort by co-located operators to corrupt or manipulate public discourse by using fake accounts and misleading people about who is behind them. They were active across the internet, including Facebook, TikTok, Twitter and Instagram, and posted memes in Malay in support of the current government coalition, with claims of corruption among its critics. On Facebook, this network managed Pages, including those posing as independent news entities, and promoted police while criticizing the opposition. Typically, their posting activity accelerated during weekdays, taking breaks for lunch. Their fake accounts were fairly under-developed and some of them used stolen profile pictures. Some of them were detected and disabled by our automated systems.

We found this network after reviewing information about a small portion of this activity initially suspected to have originated in China by researchers at Clemson University. Although the people behind it attempted to conceal their identity and coordination, our investigation found links to the Royal Malaysia Police.

- *Presence on Facebook and Instagram:* 596 Facebook accounts, 180 Pages, 11 Groups and 72 accounts on Instagram.
- *Followers:* About 427,000 accounts followed one or more of these Pages, around 4,000 accounts joined one or more of these Groups and about 15,000 accounts followed one or more of these Instagram accounts.
- *Advertising:* Around $6,000 in spending for ads on Facebook and Instagram, paid for primarily in Malaysia ringgit.

Israel

We removed 259 Facebook accounts, 42 Pages, 9 Groups and 107 Instagram accounts for violating our policy against coordinated inauthentic behavior. This network originated in Israel and targeted Angola, Nigeria and the Gaza region in Palestine.

We identified several clusters of activity that were present on multiple social media platforms and operating their own websites — with each cluster focused on a particular country. They included fictitious NGOs, media organizations and other entities that had presence across the internet, likely as "backstops" to make them appear more legitimate so they can withstand scrutiny by platforms and researchers. We found and removed the network before it was able to build its audiences, with the Nigeria-focused cluster detected shortly after it first appeared on our platform.

The individuals behind this activity relied on fake accounts — some of which were detected and disabled by our automated systems — to post, comment, manage Groups and Pages and share links to their off-platform websites. Some of these accounts posed as local independent journalists. Most of the accounts used profile pictures copied from elsewhere on the internet, while others used profile pictures likely generated using artificial intelligence techniques like generative adversarial networks (GAN). This operation appeared to have leveraged fake engagement services to buy likes in an attempt to make its content and fictitious entities appear more popular than they were, including to demonstrate the effectiveness of these campaigns to their benefactors.

This network posted primarily in English, Arabic and Portuguese about news and current events in the countries they targeted, including positive commentary about the government of Angola, one of the political candidates in Nigeria and criticism of Hamas in Gaza.

We found this network as a result of an internal investigation into the suspected coordinated inauthentic behavior in the region. Although the people behind it attempted to conceal their identities and coordination, our investigation found links to Mind Force, a PR firm in Israel. It is now banned from our platforms.

- *Presence on Facebook and Instagram:* 259 Facebook accounts, 42 Pages, 9 Groups and 107 accounts on Instagram.
- *Followers:* About 224,000 accounts followed one or more of these Pages, around 9,000 accounts joined one or more of these Groups, and about 208,000 accounts followed one or more of these Instagram accounts.
- *Advertising:* Around $12,000 in spending for ads on Facebook and Instagram, paid for primarily in Euros and US dollars.

Russia

IN-DEPTH RESEARCH & ANALYSIS:

CERTAINLY THE Z TEAM

By Mike Torrey, Security Engineer, Ben Nimmo, Global Threat Intelligence Lead, and the IO Threat Intelligence Team

EXECUTIVE SUMMARY:

We took down a network of Instagram accounts operated by a troll farm in St. Petersburg, Russia, which targeted global public discourse about the war in Ukraine. This appeared to be a poorly executed attempt, publicly coordinated via a Telegram channel, to create a perception of grassroots online support for Russia's invasion by using fake accounts to post pro-Russia comments on content by influencers and media.

We detected this activity and began taking action in March, right after reviewing public reporting by the Russian outlet Fontanka. They exposed a physical troll farm operated out of an office building in St. Petersburg, only 10 days after it had advertised job postings for "spammers, commenters, content analysts, designers and programmers" focused on YouTube, Telegram and TikTok. We took down the network in early April, once we completed our investigation, and have continued to detect and disable its attempts to come back. We linked this activity to a self-proclaimed entity called "Cyber Front Z," and to individuals associated with past activity by the Internet Research Agency (IRA). Cyber Front Z is now banned from our platforms.

Our investigation found attempts at driving comments to people's content on Instagram, Facebook, TikTok, Twitter, YouTube, LinkedIn, VKontakte and Odnoklassniki. It appears that hired "trolls" worked in shifts seven days a week, with a daily brief break for lunch. According to public reporting, they were divided into teams specializing on particular platforms they were meant to "spam." The operation had an overt and a covert component. Overtly, they ran a Telegram channel that regularly called on its followers to go to particular accounts or posts by public figures or news media and flood them with pro-Russia comments. Covertly, they ran fake accounts that posted such comments themselves — likely to make it look as if their crowdsourcing had been effective.

The targets included politicians, journalists, actors, celebrities and commercial brands from around the world — anyone who might have spoken out in support of Ukraine.

This deceptive operation was clumsy and largely ineffective — definitely not "A team" work. On Instagram, for example, more than half of these fake accounts were detected and disabled by our automated systems soon after creation. Their efforts didn't see much authentic engagement, with some comments called out as coming from trolls. We also found instances of the "trolls" who sprinkled pro-Ukraine comments on top of the paid pro-Russia commentary, in a possible attempt to undermine the operation from within.

TAKEDOWN BY THE NUMBERS

- *Presence on Facebook and Instagram:* 45 Facebook accounts and 1,037 Instagram accounts.
- *Followers:* About 49,000 accounts followed one or more of these Instagram accounts.
- *Advertising:* Around $1,400 in spending for ads on Facebook and Instagram, paid for in rubles.

THE BEANBAG TROLLS

Just like the IRA's early efforts, the Z Team was exposed by undercover journalists who responded to job ads inviting them to join the ranks of the troll farm. The Z Team story first broke in late March, when Fontanka published a long exposé of this "patriotic movement" that employed several hundred people for 45,000 roubles a month (around $440 at the time) to comment online in support of the war in Ukraine. Reportedly, the employees were divided into specific platform-focused departments — around two dozen people each — and given the login details for fake accounts to be operated from the employees' personal devices and post 200 times a day. Fontanka's photos from inside the building showed the trolls working on beanbags.

The Z Team's activity centered around the Telegram channel with over 100,000 followers as of this writing, which routinely (multiple times a day) posted a list of "targets" on many platforms, including Twitter, LinkedIn, VKontakte and Odnoklassniki. Here is one example:

Image

This public Telegram post from May 26, 2022 shares specific URLs and invites followers to comment on the Twitter and Instagram accounts of Finnish Prime Minister Sanna Marin.

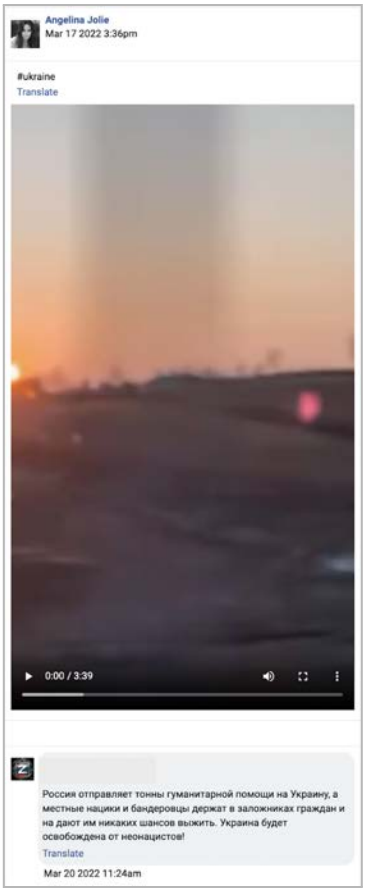
Translation

“Prime Minister of Finland Sanna Marin arrived in Ukraine. They showed her the broken towns of Irpin and Bucha, so that the Finn would cry and fork out for restoration. Of course, they blamed Russia for everything. In Kyiv, Sanna also met Zelya, who bowed and asked for military assistance and EU accession.

We must explain to the Finnish politician that Ukraine will be liberated from Nazism by the Russian army, so petitions from Zelensky from the cocaine acceptance center are not her level. 🔽 Let’s fly here and massively urge not to support the Ukronazis 🔽
In English: Stop support ukrainian nazi, Sanna! Russia will free Ukraine from the criminal regime!”

Thanks to Fontanka's reporting, we were able to build on their findings. While the original story did not mention our apps , we were able to uncover a network of related accounts on Instagram. It appears that the organizers used the people they hired as simply a typing pool to flood pro-Ukrainian posts with comments on one topic only — Russia’s war — using very basic, fake accounts that kept getting caught. They were low in sophistication, represented no distinct personas and were essentially fungible. A large portion of them were detected and disabled by automation even before we found their link to this activity. Some appeared to have been purchased from account farms around the world; others were created in batches in Russia in early March 2022. The bulk of this activity on our platform consisted of comments on other people’s content, rather than standalone posts.

Most comments were in Russian, and often replied to the substance of the posts they were commenting on, rather than reusing the same generic phrases. None of the specific comments we reviewed as part of this investigation violated our content policies. In fact, according to Fontanka, the operators were explicitly instructed not to be insulting, presumably to avoid platform enforcement. We took this network down based on its violating behavior, not the content it posted.



Image

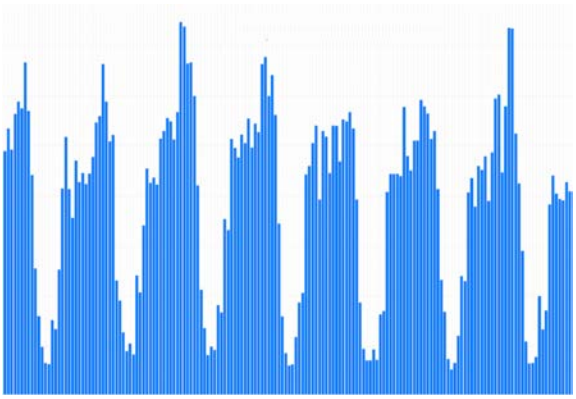
Comment under a post by Angelina Jolie from one of the Z Team's fake accounts.

Translation

"Russia is sending tons of humanitarian aid to Ukraine, and the local Nazis and Banderites are holding citizens hostage and giving them no chance to live. Ukraine will be liberated from the neo-Nazis!"

The Z Team's online behavior was similar to other troll farms we've disrupted over the years, including in Albania and Nicaragua. All of these networks posted to a fixed schedule with a clear working-day pattern, seven days a week, with a slow start in the morning and a surge toward the end of the day — possibly as the operators rushed to meet their posting quotas. The pattern showed much less of a lunchtime dip in Russia than what we observed in Albania and Nicaragua, and little variation between weekdays and the weekend.

We assess that this on-platform cluster was operated by about several dozen individuals on Instagram.



Image

Timeline of the Z Team's posting activity, Monday through Sunday.

Meeting the CIB threshold

- As part of our assessment of whether a network engages in deceptive behavior that rises to the level of CIB (rather than IB), we look at the complexity of the deceptive activity as one of the key factors.
- In certain ways, the behavior in this case was not complex. At the individual account level, we saw no serious attempts at what we call "persona-building" to make the fake accounts look believable. The network did not engage in multiple forms of influence strategies (it focused solely on mass commenting), nor did it systematically attempt to evade enforcement.
- However, the operators mostly used unique comments, with little evidence of recycling them. This shows an intentional and labor-intensive effort, distinguishing the network from other cases that fall under our spam and inauthentic engagement policies. Another feature is the extent to which the activity was targeted and tailored. The network was not sharing content in a scattershot way. Instead, the operation chose precise targets – sometimes to push back against specific pro-Ukraine messages and even memes.
- Overall, the behavior on our services was only one piece of a much larger, more complex cross-platform effort. But it is important to note that cross-platform activity alone without coordination and the unique, targeted content would not be sufficient for a CIB call, and would fall under our spam and inauthentic engagement policies.
- Finally, although the individual accounts did not engage in much persona-building, the overall operation did attempt to create the perception of a fictitious grass-roots movement, Cyber Front Z, at the cross-platform level. In reality, Cyber Front Z is a troll farm paying people to use fake profiles to deceive people about support for the war.

BACK TO THE BASICS

The people behind the Z Team took some steps to conceal their identities. However, our investigation identified links to individuals associated with past activity by the Internet Research Agency.

The Z Team's tactics closely resembled those of the IRA in its earliest days, in 2013, when it focused on targeting the Russian opposition domestically, including now-jailed activist Alexei Navalny. The earliest exposés of the IRA's Russian-language activity spoke of an office in the Olgino suburb of St. Petersburg, where teams of Russians were paid to mass-post pro-government comments on online forums, including LiveJournal. Back then, that operation advertised widely for writers, tipping off several investigative journalists, who responded to the ads, worked in the building and then exposed the now infamous troll-farm.

What we found this year is that while the Z Team updated the platforms it targeted and moved to the city center, it followed the same path as its predecessor — hiring people to mass comment and letting an undercover journalist in with them. Even the look was similar, apart from the beanbags and the "Z" flags.

Image

Open-source footage of the IRA's office in 2015, from Andrei Soshnikov/YouTube.

Image

Open-source footage of the Z Team offices in 2022, from Fontanka.

In terms of tactics, however, the Z Team behaved very differently from some of the IRA-linked operations we disrupted in 2020 and 2022. These had moved onto creating a small number of credible fake accounts that were backstopped across multiple platforms, and tried to co-opt unwitting journalists by recruiting them to work for non-existent NGOs or news outlets. Unlike the Z Team, they put a premium on deception, and relied on operators to maintain elaborate fictitious personas across platforms and withstand scrutiny while interacting with people.

FAKE v. REAL

From the start, the Z Team portrayed its mission as opposing pro-Ukrainian online activity. Its first Telegram post, on March 12, urged followers to “fight back in the information battlefield against the propagandists of the Kyiv junta funded by the Western world.”

The organizers regularly went head-to-head with pro-Ukraine organic commenting activity. Their Telegram channel identified social media posts that were receiving high volumes of pro-Ukraine comments, and directed its followers and fake accounts to respond with pro-Russian comments. The Z Team sometimes even copied memes that had been created by supporters of Ukraine, and defaced them with swastikas and other far-right imagery to promote the Russian government’s “de-Nazification” claim.

Image

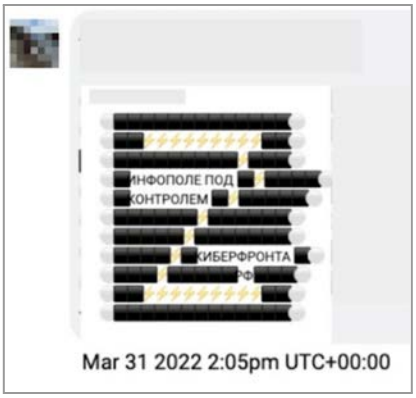
A meme created by a Ukrainian online activist group in March, highlighting the plight of civilians in the besieged city of Mariupol.

Image

The same meme copied and defaced by Cyber Front Z, showing a member of the Azov Battalion.

In many ways, the pro-Russia operation attempted to mirror the anti-war communities defending Ukraine. However, there were key differences. First, the Z Team relied on fake accounts run by paid posters. Second, anti-war and pro-Ukraine comments typically outnumbered the Z Team’s comments.

For example, on March 28, the Z Team targeted a post by the Finnish defense minister with comments that included the claim, “This info (battle) field is under the control of the Russian Federation's Cyber Front” (illustrated below). In fact, of the 255 accounts that commented on this post, almost 200 came from Ukraine, while just over 20 were operated by the Z Team.



Image

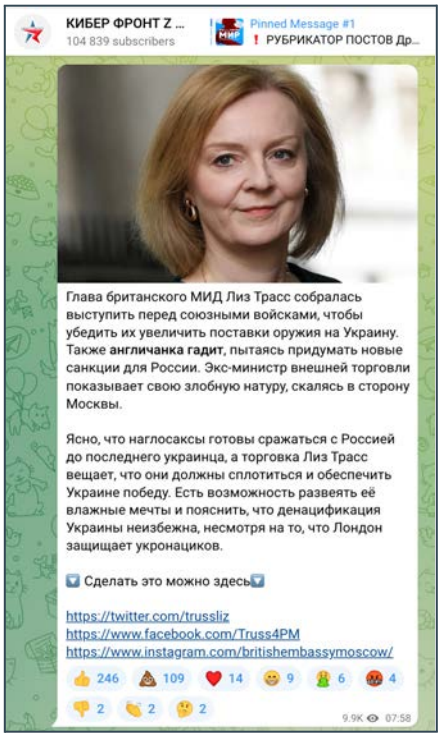
A typical comment from one of the Cyber Front Z's fake accounts, replying to a post by the Finnish minister of defense, and representing the letter "Z," characteristic of the Russian invasion.

Translation of the comment: "Info (battle)field under the control of the Russian Federation's Cyber Front."

Z-GRADE MISSES

Our investigation has not found evidence of the Z Team sparking significant support among authentic communities around the world. The regular misses by this campaign likely contributed to this lack of success.

For example, a Telegram post on May 26 invited followers to tell then-UK Foreign Secretary Liz Truss that "the de-Nazification of Ukraine is inevitable, even if London is defending the Ukro-Nazis." It listed what the Z Team thought were the relevant accounts on Twitter, Instagram and Facebook. Instead of her actual Facebook Page, however, they linked to a fan Page with about 30 followers that had not posted since 2018.

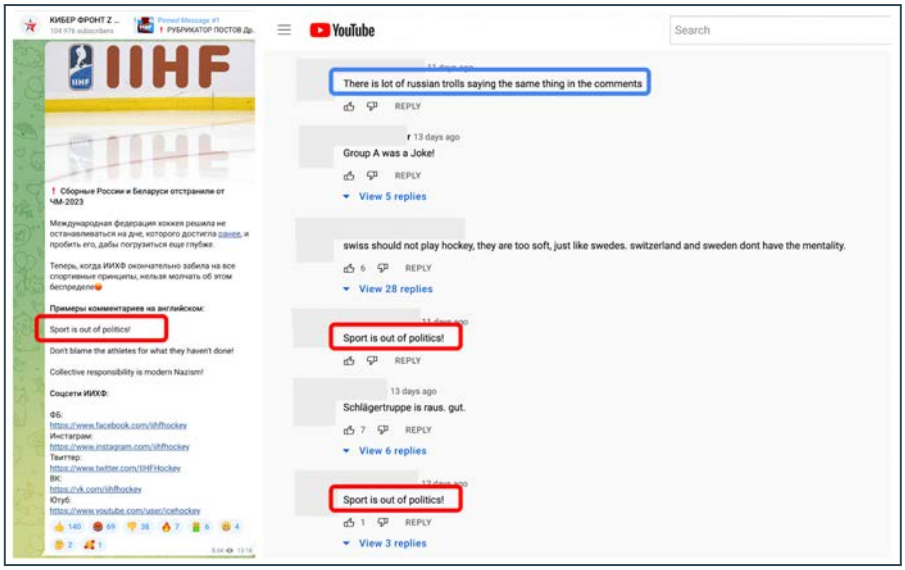


Image

Public Telegram post by the Z Team, targeting then-Foreign Secretary Liz Truss. Note the three URLs at the bottom: the Twitter link leads to her verified account, the Instagram link is that of the British Embassy in Moscow, but the Facebook Page is a fan Page unused since 2018.

In another example, the Z Team invited its followers to comment in English on the Instagram account of Finnish Prime Minister Sanna Marin. As of mid-June, these comments had not appeared anywhere on our apps. Similarly, a Telegram post on May 21 called on the Z Team's supporters to comment on the social media accounts of actor Morgan Freeman, with a similar result.

We saw more failed attempts to drum up a conversation on other platforms, including Twitter and YouTube. In late May, the Z Team steered people toward Twitter accounts including those of the President of Poland, the International Ice Hockey Federation (IIHF), the French Tennis Federation, and YouTube channels including the IIHF, Ukrainian singer Max Barskih, and Russian rock group "Машина Времени" ("Time Machine"). None of these showed a high volume of pro-Russia comments, while some people called them out as "Russian trolls."

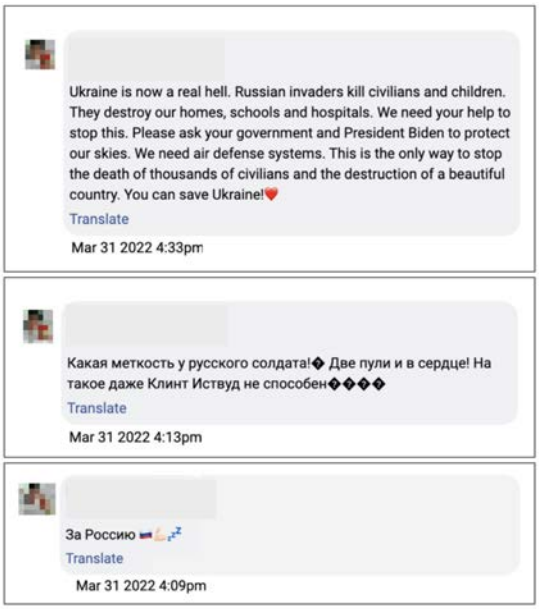


Image

Left: The Z Team providing its followers on Telegram with a set of comments for the IIHF after it banned the Russian and Belarusian teams, including the phrase "Sport is out of politics!"

Right: Comments on an IIHF YouTube video, including two quoting the Z Team message (red boxes) and another user calling out the trolls (blue box).

In isolated cases, the fake accounts appear to have assumed a split personality when posting in English versus Russian. The same account would reply to some posts with its usual pro-Russia comments, and to other posts they'd respond with pro-Ukraine comments. In some cases, they appeared to have copied and pasted pro-Ukraine comments from the very groups the Z Team explicitly opposed. This might be a case of individual operators undermining this fictitious movement from within.



Image

Three comments by the same fake Instagram account on March 31, 2022, in the span of 24 minutes. The account posted the English text, which appeared to have been copied and pasted from other people's pro-Ukraine posts, four times in three minutes.

Translation

How accurate the Russian soldier is! Two bullets through the heart! Not even Clint Eastwood could do that For Russia 🇷🇺 💪 💤

These examples underscore the importance of analyzing attempted influence operations according to the evidence, and not taking any claims of viral success at face value. Some threat actors try to capitalize on the public's fear of influence operations by trying to create the false perception of widespread manipulation, even if there is no evidence — a phenomenon we called out in 2020 as "perception hacking." It is possible that operators may also do this in an attempt to convince their funders or employers of their effectiveness in corralling large-scale authentic movements, while "faking" this engagement on the back end. However, the available evidence suggests that they haven't succeeded in rallying substantial authentic support online as part of this operation. Interestingly, the Z Team's activity was publicized as a "patriotic movement" by a number of Russian media entities, including those previously reported to have links with the IRA. This amplification aspect offers an opportunity for future open-source research.

Appendix: Threat indicators

1. BITTER APT

Domains & C2s

Domain	Description
signalpro[.]org	Hosting Dracarys Malware
signal-premium[.]org	Hosting Dracarys Malware
signalpremium[.]com	Hosting Dracarys Malware
telegram-pro[.]org	Hosting Dracarys Malware
signal-premium-app[.]org	Dracarys Malware C2
youtubepremiumapp[.]com	Dracarys Malware C2
pflix.camdvr[.]org	Dracarys Malware C2
94.140.114[.]22	Dracarys Malware C2
weather.play-protect[.]com	Assessed to be attacker controlled infrastructure
gallery.play-protect[.]com	Assessed to be attacker controlled infrastructure
sikhsiyasatapp[.]net	Assessed to be attacker controlled infrastructure
telegramapppro[.]org	Assessed to be attacker controlled infrastructure
play-protect[.]com	Assessed to be attacker controlled infrastructure
www.sikhsiyasatapp[.]net	Assessed to be attacker controlled infrastructure
briarapppro[.]org	Assessed to be attacker controlled infrastructure
islam-360-plus[.]com	Assessed to be attacker controlled infrastructure
converse-app[.]org	Assessed to be attacker controlled infrastructure
telegram-app[.]tech	Assessed to be attacker controlled infrastructure
appprotonvpn[.]com	Assessed to be attacker controlled infrastructure

linphone-app[.]com	Assessed to be attacker controlled infrastructure
appbriar[.]com	Assessed to be attacker controlled infrastructure
gosignal[.]org	Assessed to be attacker controlled infrastructure
app2.appvlc[.]com	Assessed to be attacker controlled infrastructure

Hashes

MD5	Description	Malware Family
a3d18021cd444e8fe23fffc1a6140071	Signal Pro	Dracarys
07532dea34c87ea2c91d2e035ed5dc87	Youtube Premium	Dracarys
e20473bea7fe5968f0a032303838b601	Signal Pro	Dracarys
d9a39c41e9f599766b5527986e807840	pflix	Dracarys
b06e2f95ecf7012138bee314be9baed9	pflix	Dracarys

2. APT36

Domains & C2s

Domain	Description
1drivestorage[.]com	Assessed to be actor-controlled domain hosting malware
appsupdate[.]net	Assessed to be actor-controlled domain hosting malware
archiverst[.]com	Assessed to be actor-controlled domain used to redirect to other actor-controlled domains
filestudios[.]net	Assessed to be actor-controlled domain hosting malware
hatvax[.]com	C2 for malware
medizz[.]co	C2 for malware
play[.]google[.]com[.]whatsapp[.]playapps[.]ga	Assessed to be actor-controlled domain hosting malware
ratapi11223344786[.]azurewebsites[.]net	C2 for malware
rdeskapi719543132892786[.]azurewebsites[.]net	C2 for malware
rkarsin453287786[.]azurewebsites[.]net	C2 for malware
secureapplication[.]azurewebsites[.]net	C2 for malware
securechat[.]azurewebsites[.]net	C2 for malware

shareflx[.]com	Assessed to be actor-controlled domain hosting malware
shareflx[.]createasocialcard[.]top	Social card preview site that redirects to actor-controlled domain
shareflx[.]social-card-share[.]top	Social card preview site that redirects to actor-controlled domain
shareflx[.]socialpreviews[.]top	Social card preview site that redirects to actor-controlled domain
storeupdates[.]net	Assessed to be actor-controlled domain hosting malware
testandroidopen[.]azurewebsites[.]net	C2 for malware
theambix[.]org	C2 for malware
yoursdrive[.]com	Assessed to be actor-controlled domain hosting malware

Hashes

- `5d885fd9b896c8d59dbdc6b3ae4068662544f401d98a7eba757b329714d87c45`
- `b3510e0a8775d9ab5c8409510041dc1e7da47923d5bf3e8f0848a4a3970ffca7`
- `7999f5af42e6a825db56aa800a6b957c19d609225cc339f12cf85dde06af3b74`
- `5d9027c76306efd5fb57f42dbbaa26f976657a523c32d8fd3fa628ee1417d0aa`

Yara rule

```
rule xploitspy_rat {

  meta:
```

```
    source = "Facebook"

    date = "2022-08-04"

    description = "Android RAT found on GitHub at
https://github.com/XploitWizer/XploitSPY/tree/master/client/app/src/main/java/
com/remote/app."

  strings:

    $func0 = "0xAU"

    $func1 = "0xCL"

    $func2 = "0xCO"

    $func3 = "0xFI"

    $func4 = "0xGP"

    $func5 = "0xIN"

    $func6 = "0xLO"

    $func7 = "0xMI"

    $func8 = "0xPM"

    $func9 = "0xSM"

    $func10 = "0xWI"

    $func11 = "0xCB"

    $func12 = "0xNO"

    $applist0 = "appName"

    $applist1 = "packageName"

    $applist2 = "versionName"

    $applist3 = "versionCode"

    $notif0 = "appName"

    $notif1 = "postTime"

  condition:
```

```
    7 of ($func*) and (
      all of ($applist*)
      or all of ($notif*)
    )
}

rule lazaspy_android_rat {
  meta:
    source = "Facebook"
    date = "2022-08-04"
    description = "Custom Android RAT built on top of XploitSPY"
  strings:
    $s0 = "/.System/Ct.csv/"
    $s1 = "/.System/sm.csv/"
    $s2 = "logg.txt"
    $s3 = "ulog.txt"
    $s4 = "This Feature is currently Unavailable. Comming Soon!"
    $s5 = "Press Back Again to Exit."
    $s6 = "Please Grant Permission to Continue"
    $s7 = "Try Again something went wrong"
    $s8 = "Deleting Conversation Please wait"
    $s9 = "please type something"
    $s10 = "Message not Sent"
  condition:
    7 of ($s*) and xploitspy_rat
}
```