WhatsApp

Stopping Abuse:

How WhatsApp Fights Bulk Messaging and Automated Behavior

February 6, 2019

Contents

Summary

From the early days of email to today’s social networks, internet platforms have long had to confront abuse including bulk messaging and automated behavior that spreads problematic content. WhatsApp takes a multi-pronged approach to this challenge while also working to maintain the private nature of our service. We have built sophisticated machine learning systems to detect abusive behavior and ban suspicious accounts at registration, during messaging, and in response to user reports. We remove over two million accounts per month for bulk or automated behavior — over 75% without a recent user report. These efforts are particularly important during elections where certain groups may attempt to send messages at scale. While there are many actors trying to abuse the free service we provide, we are constantly advancing our anti-abuse operations to keep our platform safe.

Introduction

WhatsApp was built for private messages: to help people chat with their loved ones, conduct business, or talk confidentially with a doctor. Instead of encouraging users to build an audience and share widely, WhatsApp is designed to help people share with others they know or get helpful information from a business. To protect the privacy and security of our users, WhatsApp provides end-to-end encryption by default, which means only the sender and recipient can see the content of messages.

Our service is not a broadcast platform. We place limits on group sizes and how users send messages. Approximately 90% of the messages sent on WhatsApp are from one person to another, and the majority of groups have fewer than ten people. WhatsApp requires the message sender to know the phone number of the recipient. Whenever a user receives a message from someone not in their address book, we ask the user if they want to block or report the sender.

As with any communications platform, sometimes people attempt to exploit our service. Some may want to distribute click-bait links designed to capture personal information, while others want to promote an idea. Regardless of the intent, automated and bulk messaging violates our terms of service and one of our priorities is to prevent and stop this kind of abuse.

This paper explains our approach. We have had to balance our desire for transparency while not undermining the effectiveness of these anti-abuse measures. Given the nature of end-to-end encryption, we focus on the behavior of accounts and user-reported content[1]. We employ a team of engineers, data scientists, analysts, and researchers with a wide range of backgrounds and expertise to deter and prevent abuse. We pay close attention to user feedback and engage with experts in misinformation, cybersecurity, and election integrity. Through this comprehensive effort we seek to make it difficult for adversaries to abuse the platform we created to connect people.

[1] WhatsApp scans unencrypted information like profile photos for child exploitative imagery. Please see this FAQ for more.

WhatsApp Design

While many online services require an email to sign up, WhatsApp is based on phone numbers and requires that all users have one. This presents unique opportunities for us to catch those seeking to abuse our service.

Mobile numbers are activated through SIM cards, which are distributed by carriers. Buying a SIM card is often simple for a normal user, but SIM cards can be logistically difficult to acquire at scale. In some places, identification is required for purchase; costs may vary across geographies. We design our abuse detection with these factors in mind. Phone numbers originating from areas with a history of fraud may indicate to WhatsApp there is a problem with an account. It can also be a signal to users because receiving a message from an unfamiliar number often indicates something is not right.

Of course, other types of messaging services see abuse as well. A recent analysis of U.S. communications services cited by the CTIA indicated that 2.8% of SMS messages and 50% of e-mail are spam. As on these platforms, preventing users from receiving these kinds of messages is a top priority. Since turning on end-to-end encryption by default in 2016, WhatsApp has worked to reduce spam to maintain the health of our service.

Sending mass messages from a mobile phone is time consuming. It's why people attempting to send bulk messages often build hardware or software to try and automate the process. However, those steps pose financial cost and risk. Our technology aims to identify accounts using automation and take action to stop them. In so doing, we aim to raise the costs associated with this abuse and deter bad actors from trying in the future.

Detecting and Tackling Abuse

Attempting to distribute content en masse on WhatsApp requires users to work around the design of our platform. Doing so exhibits signals that we use to identify abusive accounts. When we are confident that an account is abusive, we ban the account from WhatsApp altogether. We then use the information available to us to reverse engineer past behavior to prevent similar abuse in the future.

Our goal is to identify and stop abusive accounts as quickly as possible, which is why identifying these accounts manually is not realistic. Instead, we have advanced machine learning systems that take action to ban accounts, 24 hours a day, 7 days a week.

This abuse detection operates at three stages of an account's lifestyle: at registration; during messaging; and in response to negative feedback, which we receive in the form of user reports and blocks. A team of analysts augments these systems to evaluate edge cases and help improve our effectiveness over time. We continually work to detect abuse as early as possible to have the greatest impact at preventing abuse before it can reach people.

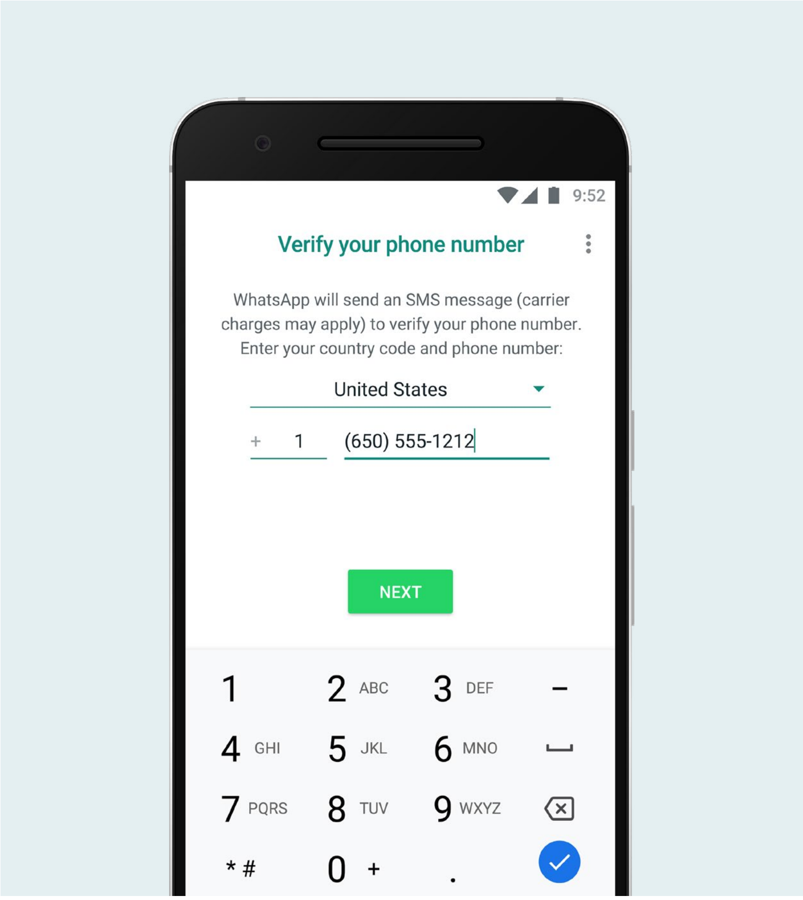
Over the last three months, WhatsApp banned over two million accounts per month for bulk or automated behavior and over 75% of those accounts did not have any recent user reports. To address rare cases when we ban an account in error, we also have a team that reviews and responds to user appeals. In that process, users may be asked for additional information in order to regain access to their account.

2 million
Accounts banned monthly

1.5 billion
Monthly active users

At Registration

All accounts must pass through a common checkpoint: registration. This is the stage where WhatsApp sends a temporary code via SMS or a phone call. Users must enter this one-time code to prove they have control over the phone number and SIM card for their account. We also provide optional two-step verification to prevent attackers from attempting to take over accounts they do not actually own. Since all accounts must go through registration before being able to send any messages, it is an effective starting point to concentrate our efforts.



Verify your phone number

WhatsApp will send an SMS message (carrier charges may apply) to verify your phone number. Enter your country code and phone number:

United States

+ 1 (650) 555-1212

WhatsApp cares deeply about the privacy of our users, which is why we do not ask for much information from them. Yet even basic account information along with an IP address and associated carrier information can be used to teach our machine learning systems the difference between bulk and normal registrations.

Because we ban accounts that send a high volume of messages, coordinated campaigns often try to spread their activity across many different accounts. We therefore work to understand the behavioral cues indicating bulk registrations. For example, our systems can detect if a similar phone number has been recently abused or if the computer network used for registration has been associated with suspicious behavior. As a result, we're able to detect and ban many accounts before they register — preventing them from sending a single message. In the same three month period, roughly 20% of account bans happened at registration time.

over 75%

accounts banned without a recent user report

roughly 20%

accounts banned at registration

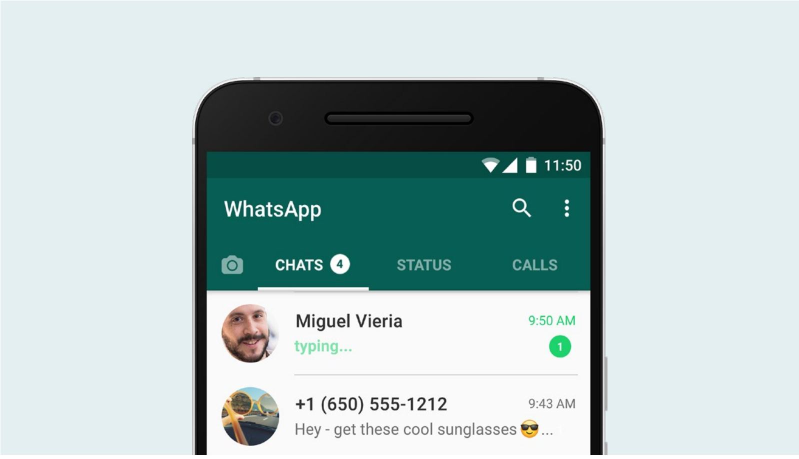
While Messaging

For accounts that complete registration, we evaluate how they behave in real time. All messages are end-to-end encrypted, which means that WhatsApp cannot see the content of messages passing through our system, though we are able to analyze the frequency of account activity.

Normal users operate relatively slowly on WhatsApp, tapping messages one at a time or occasionally forwarding content. The intensity of user activity can provide a signal that accounts are abusing WhatsApp.

For example, an account that registered five minutes before attempting to send 100 messages in 15 seconds is almost certain to be engaged in abuse, as is an account that attempts to quickly create dozens of groups or add thousands of users to a series of existing groups. We ban these accounts immediately and automatically. In less-obvious situations, a new account might message dozens of recipients who do not have the sender's account in their contacts. This could be the beginning of a spam attack, or it could be an innocent user simply telling their contacts about a new phone number. In these cases, we consider historical information (for instance, how suspicious their registration was) in order to separate abnormal — but innocuous — user behavior. In sum, our detection systems evaluate hundreds of factors to shut down abuse.

We also build detection mechanisms that directly target automation. For example, we display at the top of a chat thread when a user is typing. Spammers attempting to automate messaging may lack the technical ability to forge this typing indicator. If an account continually sends messages without triggering the typing indicator, it can be a signal of abuse, and we will ban the account. This is just one example of the combinations of actions or inactions that our artificial intelligence automatically analyzes to make decisions on which accounts to ban.



Miguel Vieria 9:50 AM
typing... 1

+1 (650) 555-1212 9:43 AM
Hey - get these cool sunglasses 😎 ...

We have advanced our technology to the point where we combine signals to make rapid determinations without human intervention. We learn from how other users have behaved and surmise the reputation of new accounts. For example, if an active computer network has recently been used by known abusers, we have more reason to believe a new account on that network is likely to be abusive. These behaviors, and other signals about current and past behavior, are helpful in finding coordinated campaigns that are trying to abuse WhatsApp.

It is important that we try and understand the diverse motivations at hand. For example, someone that has an economic incentive may want to encourage unsuspecting users to tap on a suspicious link, which could lead to a website that harvests personal information from a user. The fact we ban accounts quickly drives up their costs so that any incremental revenue gained through such tactics may not be worth the expense to try in the first place.

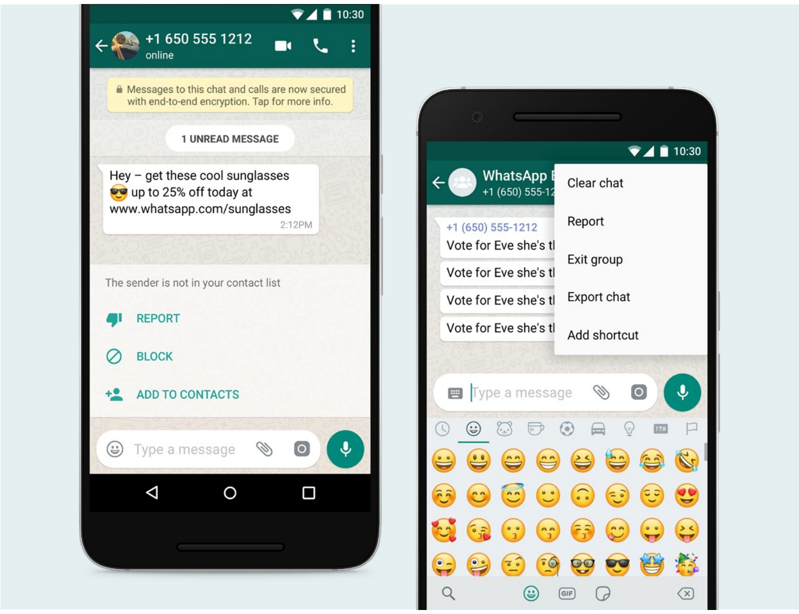
The goals of someone with a political motivation, such as attempting to distort the public discussion with false information, could be different. They seek to organize and drive interest within a particular country, region, or even city - often within a limited time window. To do so they may acquire targeted lists of phone numbers from third parties to message users over WhatsApp without their consent. For this reason, we maintain limits on how many groups an account can create within a certain time period and ban accounts with suspicious group behavior, if appropriate, even if their activity rates are low or have yet to demonstrate high reach. Whenever a user gets added to a group from someone outside their contact list we display an option that asks if the user wants to “report” or “exit” the group. Given the increase concern about abusive activities from politically motivated actors, this is an area of deep focus for our team.

In Response to Negative Feedback

Finally, if an account accumulates negative feedback, such as when other users submit reports or block the account, our systems evaluate the account and take appropriate action.

WhatsApp makes it easy for users to report a problem, and we encourage users to do so. Whenever a user receives a message for the first time from an unknown number, we immediately display options that enable them to “report” or “block” the sender’s account. In addition, users can report a problem to WhatsApp at any time in an individual or group chat.

When a user sends a report, our machine learning systems review and categorize the reports so we can better understand the motivation of the account, such as whether they are trying to sell a product or seed misinformation. This new information not only helps us ban accounts, but it also helps us improve our machine learning systems to take earlier action in the future (such as at registration or while messaging, as described above).



Messages to this chat and calls are now secured with end-to-end encryption. Tap for more info.

Hey – get these cool sunglasses 😎 up to 25% off today at www.whatsapp.com/sunglasses
2:12PM

The sender is not in your contact list

+1 (650) 555-1212
Vote for Eve she's tl
Vote for Eve she's tl
Vote for Eve she's tl
Vote for Eve she's tl

While WhatsApp seeks to quickly ban abuse, we recognize the reporting tool can be manipulated. Suppose one group of actors wants to prevent an individual from using WhatsApp. They might organize other users to make a series of reports against their victim. If our systems looked only at report rates and banned any accounts exceeding a threshold, attacks like this could knock legitimate users off our platform. Instead, we evaluate the relationship between the reporter and the reported user to rapidly evaluate the behavior. For instance, when someone reports the first message they receive from an unknown number, we have higher confidence that the report is legitimate. In cases when the reported number did not initiate the communication, we work to ensure there was no coordination among others to falsely report the account.

Leading Advances with Machine Learning

We are constantly working to stay ahead of highly-motivated abusers. As WhatsApp has become more popular we have expanded our detection efforts. The signals we analyze have become complex enough that they're too numerous for an individual to evaluate on a case-by-case basis. Fortunately, computers are very good at considering a large number of factors at once. To that end, WhatsApp has developed machine learning models to help spot abuse quickly.

Machine learning models require basic components known as features, labels, and infrastructure to complete their tasks. Features are the specific signals — such as number of reports, rate of messaging, reputation of other users sharing the same computer network, and so on — that provide insight into abusive behavior. We use labels to mark the worst offenders and distinguish between their behaviors from those of regular users. We use the features and labels to teach our systems to better predict whether a user is likely to be banned in the future. For example, if they determine that an account's behavior at registration matches the behavior of other accounts that were banned, we will ban the account even before it can send any messages. If these systems previously caught accounts after they sent 50 messages, a machine learning model trained on the features and labels from those accounts will catch similar attempts much faster.

Finally, we need infrastructure to train these classifiers, evaluate their performance, and analyze the sending behavior of actual users. Here we worked to adapt the so-called "Facebook Immune System" for use on social media to our messaging service.

In the future, we want to increase the speed of these systems by ensuring they can get real-time labels of abusive accounts they have missed and good accounts they incorrectly identified as abusive. This will help our machine learning models adapt and respond immediately to any new kind of abuse. The more we can increase the cost for those who attempt to abuse our platform, the less they will succeed.

Addressing Challenges Beyond Our Product

We have also seen that some attackers attempt to modify WhatsApp software and trade unauthorized APK files to get around the constraints coded into the client itself. These modifications cannot circumvent our detection systems that run on our servers and apply to all users without exception. Still, we are improving our ability to detect these modified versions of WhatsApp as they both violate our Terms of Service and pose a security risk to users that have installed them on their device. Users should only download WhatsApp from our website or an App Store like Google Play or the Apple App Store.

We also go beyond the platform where possible. We work with providers to ban third-party apps that claim to provide unauthorized services for our platform. In addition, when we see companies claiming they can send bulk messages on WhatsApp, we demand they cease operations and reserve all rights available to us under applicable law.

Privacy and Safety

We strongly believe that providing a platform for private communication contributes to the safety of our users. People expect their messages to be secured so they do not fall into the wrong hands. In some places WhatsApp provides a lifeline to human rights organizations and journalists. Through a combination of efforts, we can help prevent abuse and continue to provide the privacy and security benefits that come with end-to-end encryption.

Fighting automated and bulk messaging is just one way that we help users stay safe. We also modify the way our product is designed based on how it is used. We now label forwarded messages to notify users when the message they received was not created by the sender. We have limited how many forwarded messages can be sent at once. In several countries we have launched public education campaigns to help users spot hoaxes and caution them to not spread rumors.

We take a broad view of our responsibility to assist users and this comprehensive approach helps maintain privacy and safety on WhatsApp. We encourage users who receive unwanted messages to block and report accounts so we can continue to improve WhatsApp and provide a valuable service around the world.