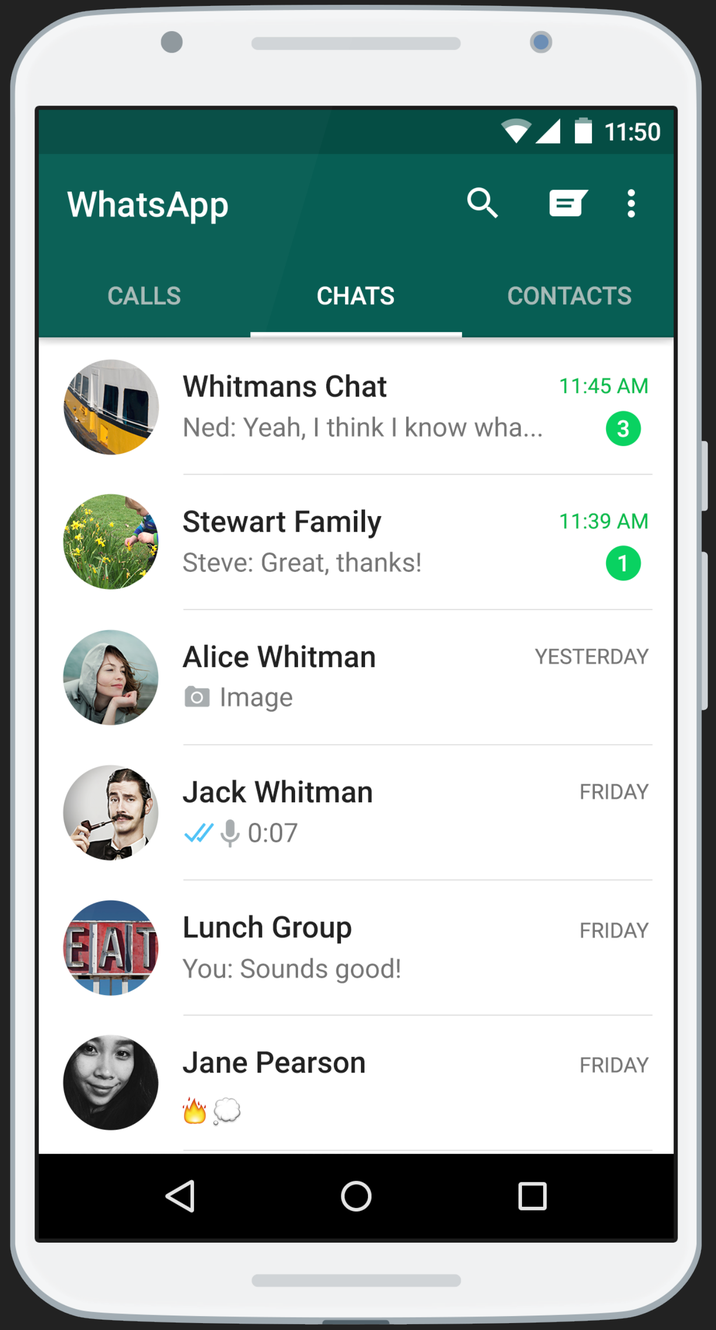
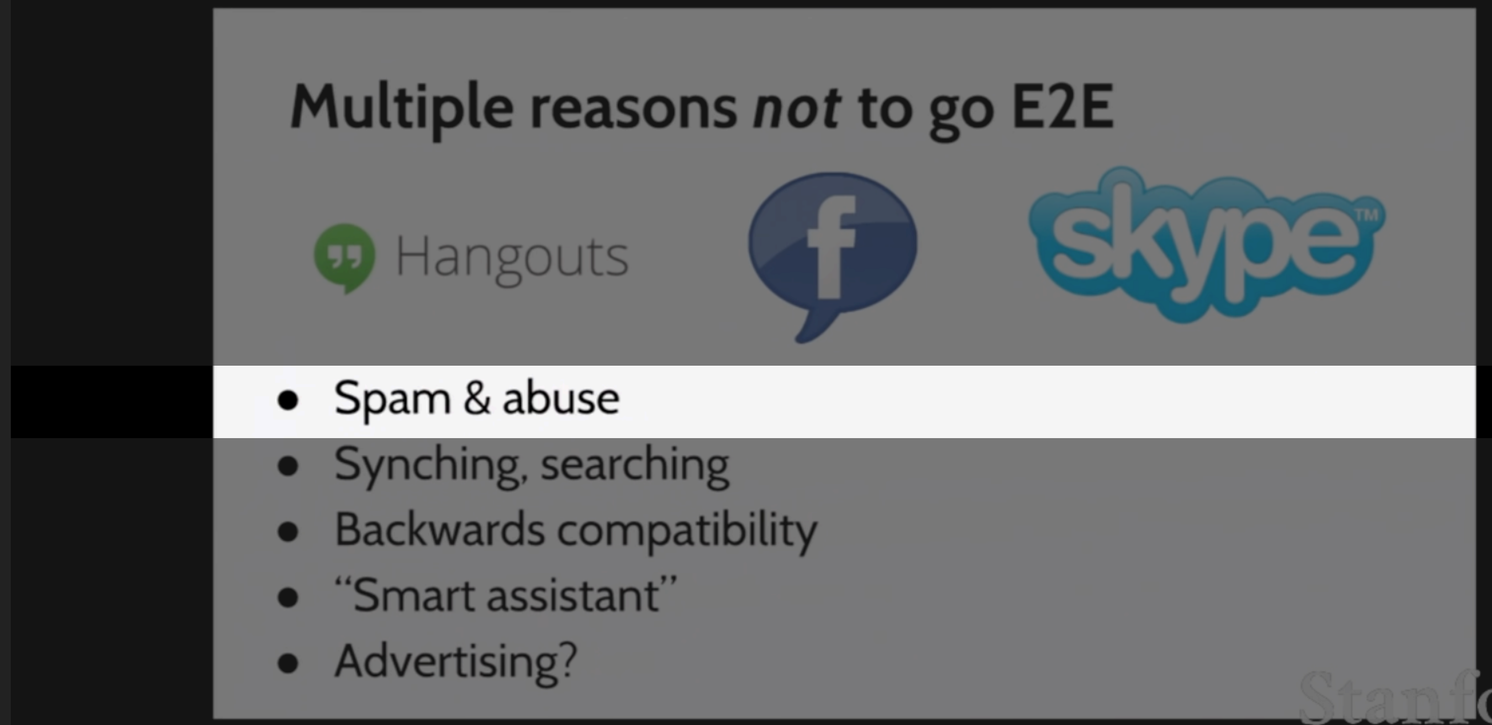
MATT JONES

HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING END-TO-END ENCRYPTION


11:50
WhatsApp
CALLS
CHATS
CONTACTS
Whitmans Chat
11:45 AM
Ned: Yeah, I think I know wha...
3
Stewart Family
11:39 AM
Steve: Great, thanks!
1
Alice Whitman
YESTERDAY
Image
Jack Whitman
FRIDAY
0:07
Lunch Group
FRIDAY
You: Sounds good!
Jane Pearson
FRIDAY

MATT JONES

HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING END-TO-END ENCRYPTION



Stanford Seminar - Challenges in Secure Messaging

stanfordonline

Subscribe 41,789

803 views

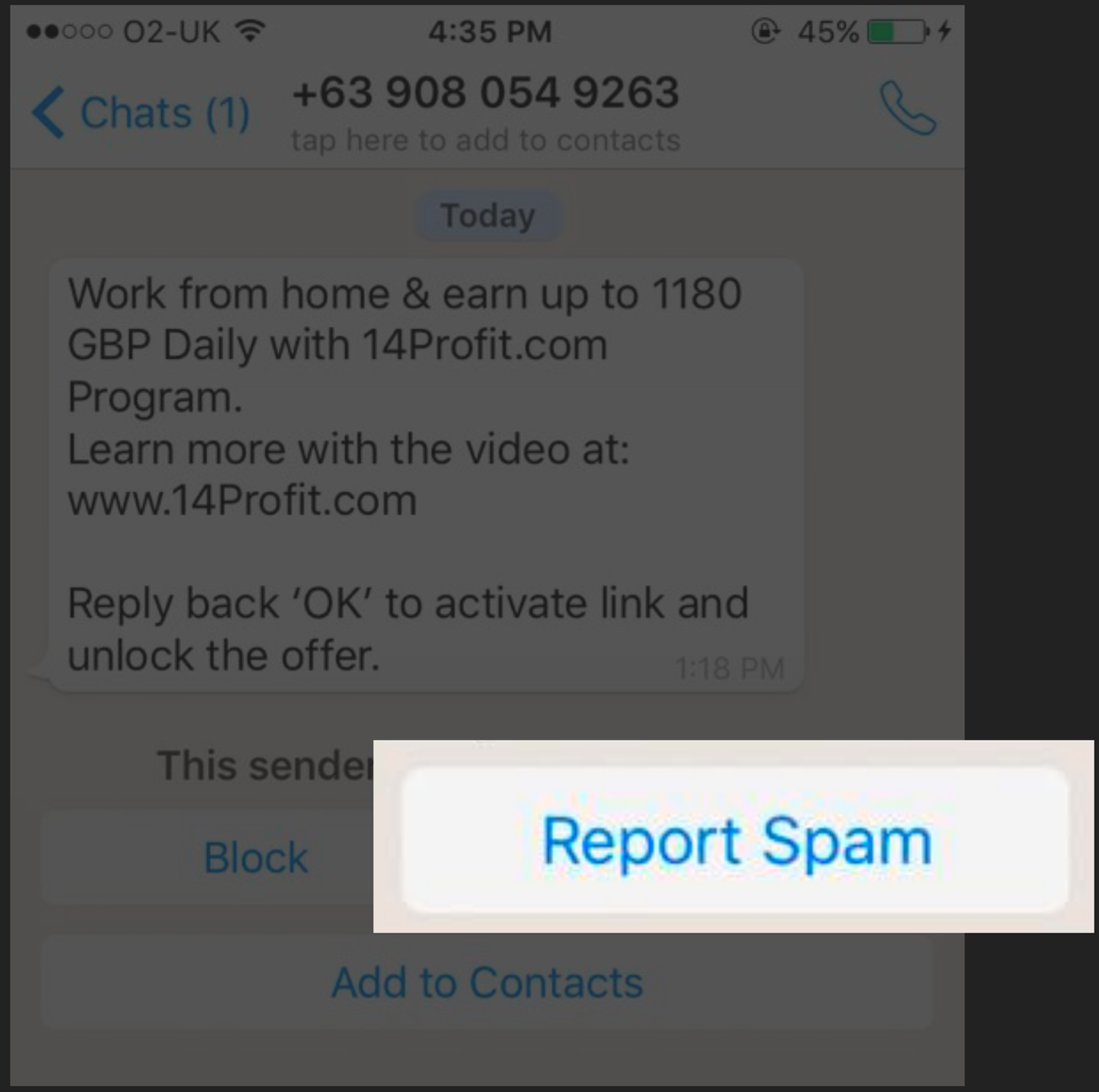
Add to Share More

16 0

Published on Dec 1, 2016

"Challenges in secure messaging" - Joseph Bonneau, Stanford University and EFF

Support for the Stanford Colloquium on Computer Systems Seminar Series provided by the Stanford Computer Forum.

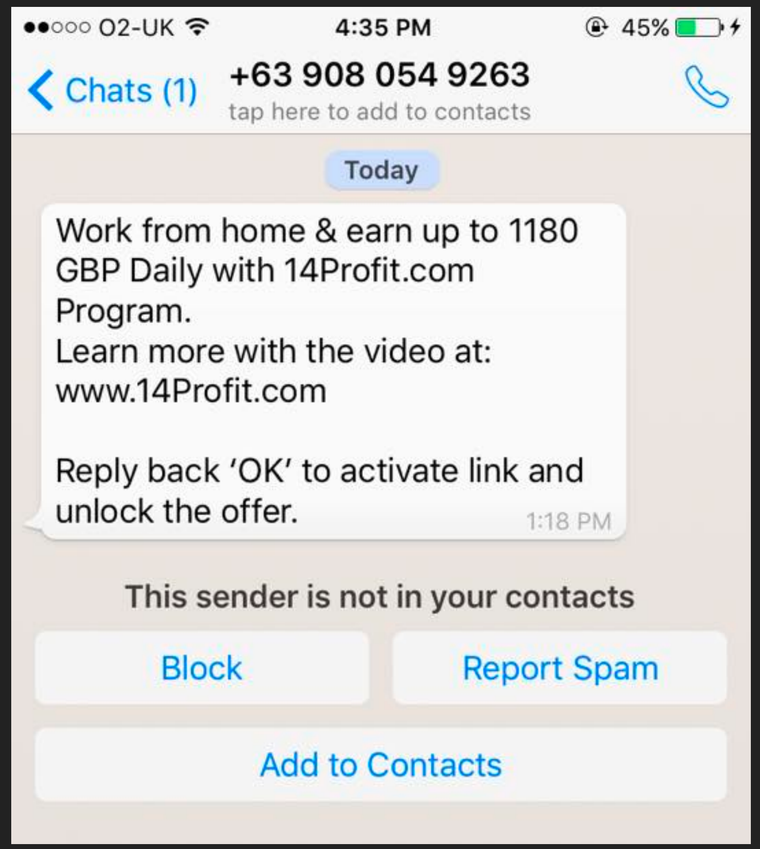

O2-UK
4:35 PM
45%
Chats (1)
+63 908 054 9263
tap here to add to contacts
Today
Work from home & earn up to 1180 GBP Daily with 14Profit.com Program.
Learn more with the video at: www.14Profit.com
Reply back 'OK' to activate link and unlock the offer.
1:18 PM
This sender
Block
Report Spam
Add to Contacts

MATT JONES

HOW WHATSAPP REDUCED SPAM WHILE LAUNCHING END-TO-END ENCRYPTION

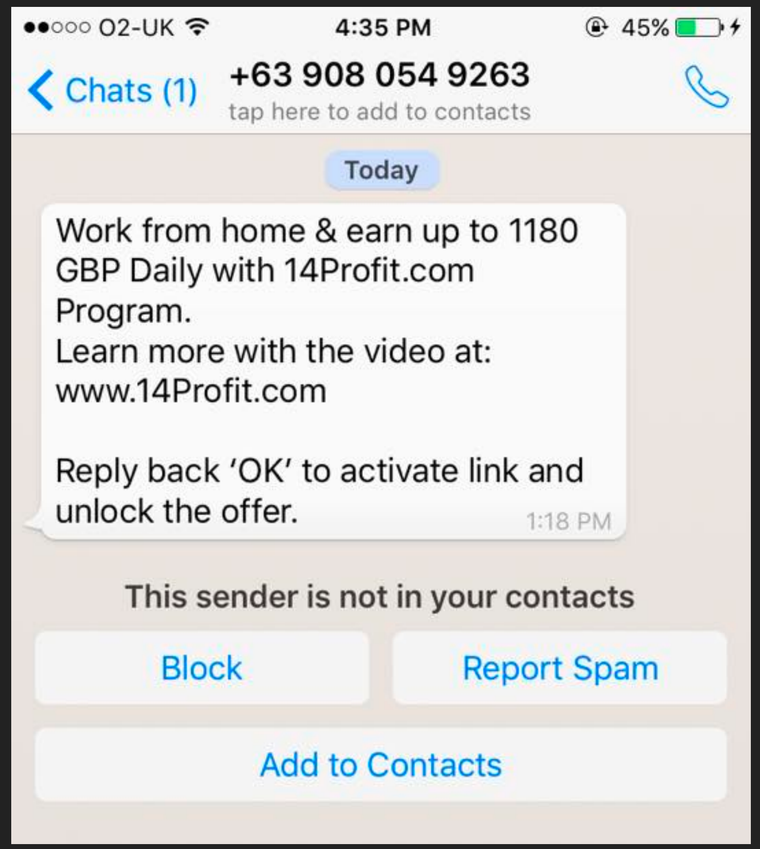
SPAM ON WHATSAPP

- Unwanted
- Commercial
- High-volume
- Automated
- From Malicious Accounts



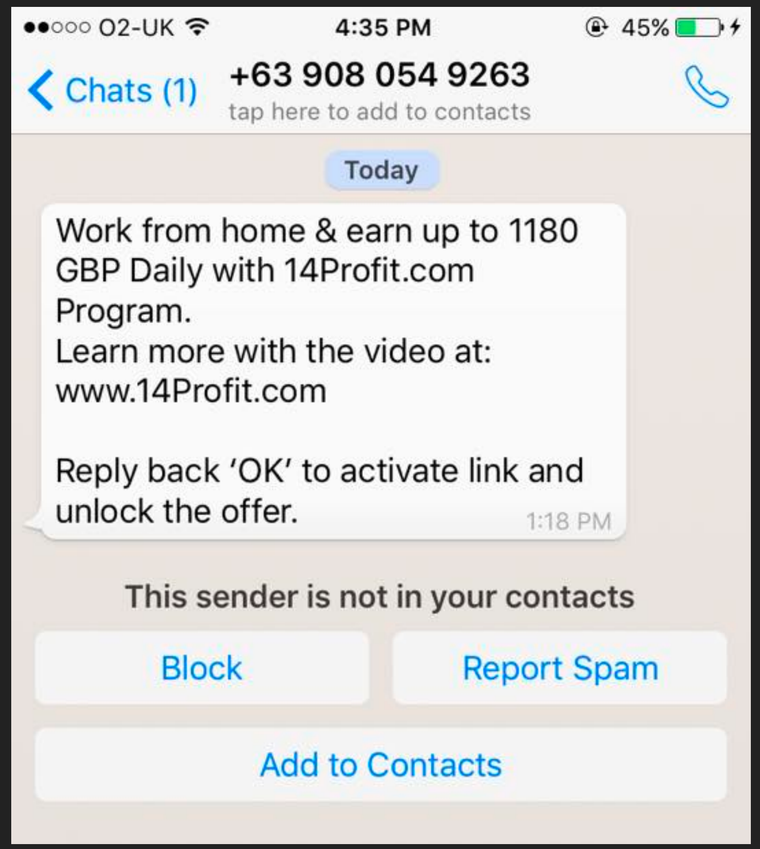
SPAM ON WHATSAPP

- Unwanted
- Commercial
- High-volume
- Automated
- From Malicious Accounts



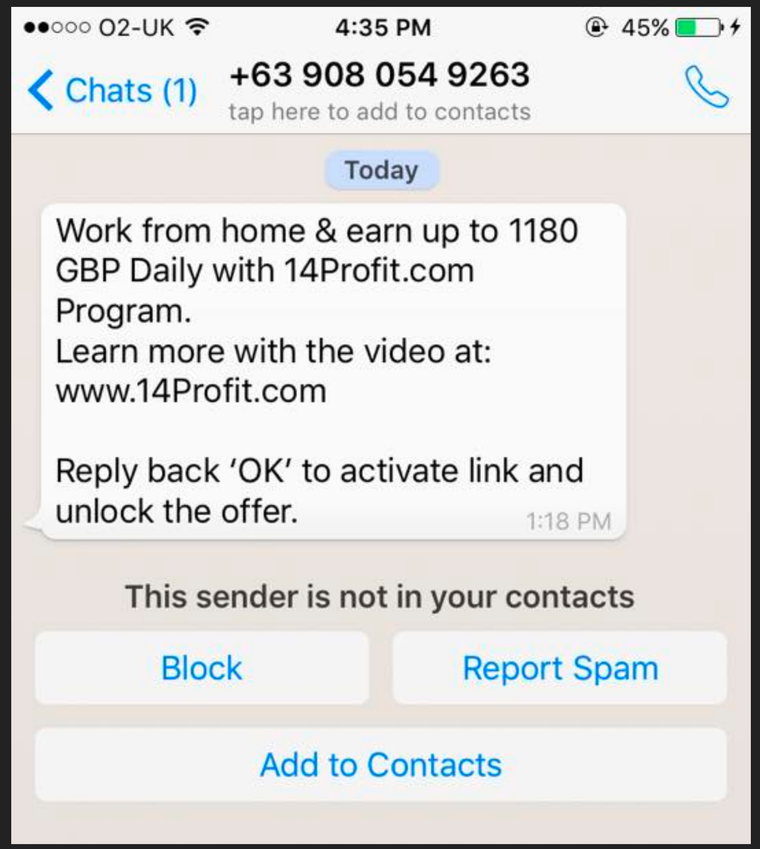
SPAM ON WHATSAPP

- Unwanted
- Commercial
- High-volume
- Automated
- From Malicious Accounts



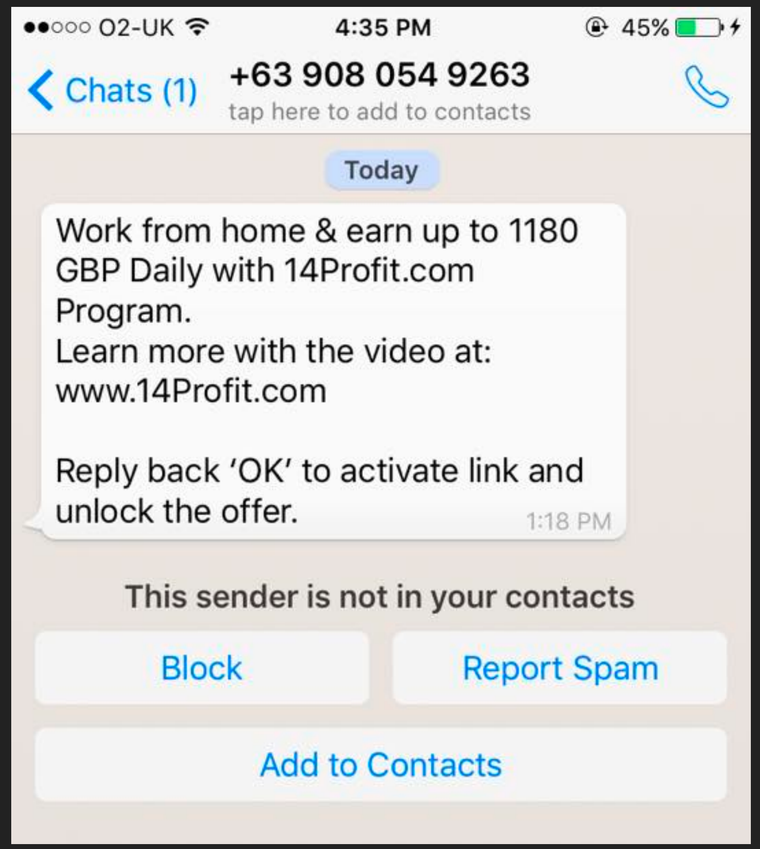
SPAM ON WHATSAPP

- Unwanted
- Commercial
- High-volume
- Automated
- From Malicious Accounts



SPAM ON WHATSAPP

- Unwanted
- Commercial
- High-volume
- Automated
- From Malicious Accounts



REDUCE TO EASIER PROBLEM

- Spam?
  - Spam account? (No passwords = no ATO)
    - Automated? (High-volume = automated)
      - Spam client? (automated = script or emulator)

Faster

Less effective

GENERAL WORKFLOW



GENERAL WORKFLOW



TRAINING DATA

Features	Label
user age=5280 days, # messages 30sec=0…	Ham
user age=2 minutes, # messages 30sec=10…	Spam
user age=3 months, # messages 30sec=2…	Ham
…	…

TRAINING DATA

Features	Label
user age=5280 days, # messages 30sec=0...	???
user age=2 minutes, # messages 30sec=10...	???
user age=3 months, # messages 30sec=2...	???
...	

TRAINING LABELS

- Need source of truth for labels in classifier training
- 2 mechanisms:
  - People we eventually caught
  - “Spam App” detection

TRAINING LABELS

- Need source of truth for labels in classifier training
- 2 mechanisms:
  - People we eventually caught
  - “Spam App” detection

TRAINING LABELS

- Need source of truth for labels in classifier training
- 2 mechanisms:
  - People we eventually caught
  - “Spam App” detection

mgp25 / Chat-API

Watch 466 | Star 2,714 | Fork 1,341

Code | Issues 1 | Pull requests 0 | Projects 0 | Wiki | Pulse | Graphs

The php WhatsApp library

679 commits | 1 branch | 14 releases | 55 contributors | GPL-3.0

Branch: master | New pull request | Find file | Clone or download

mgp25 committed on GitHub Merge pull request #2132 from arnoldnijboer/master ... Latest commit 5004d19 on Sep 6

.github	Update ISSUE_TEMPLATE.md	8 months ago
examples	updated whatsapp.php	3 months ago
src	Update Registration.php	a month ago
tests	Missed a file to include	3 months ago
.gitignore	Encryption support	a year ago
LICENSE	Encryption support	a year ago
README.md	Update README URLs based on HTTP redirects	9 months ago
composer.json	Composer updates	8 months ago

README.md

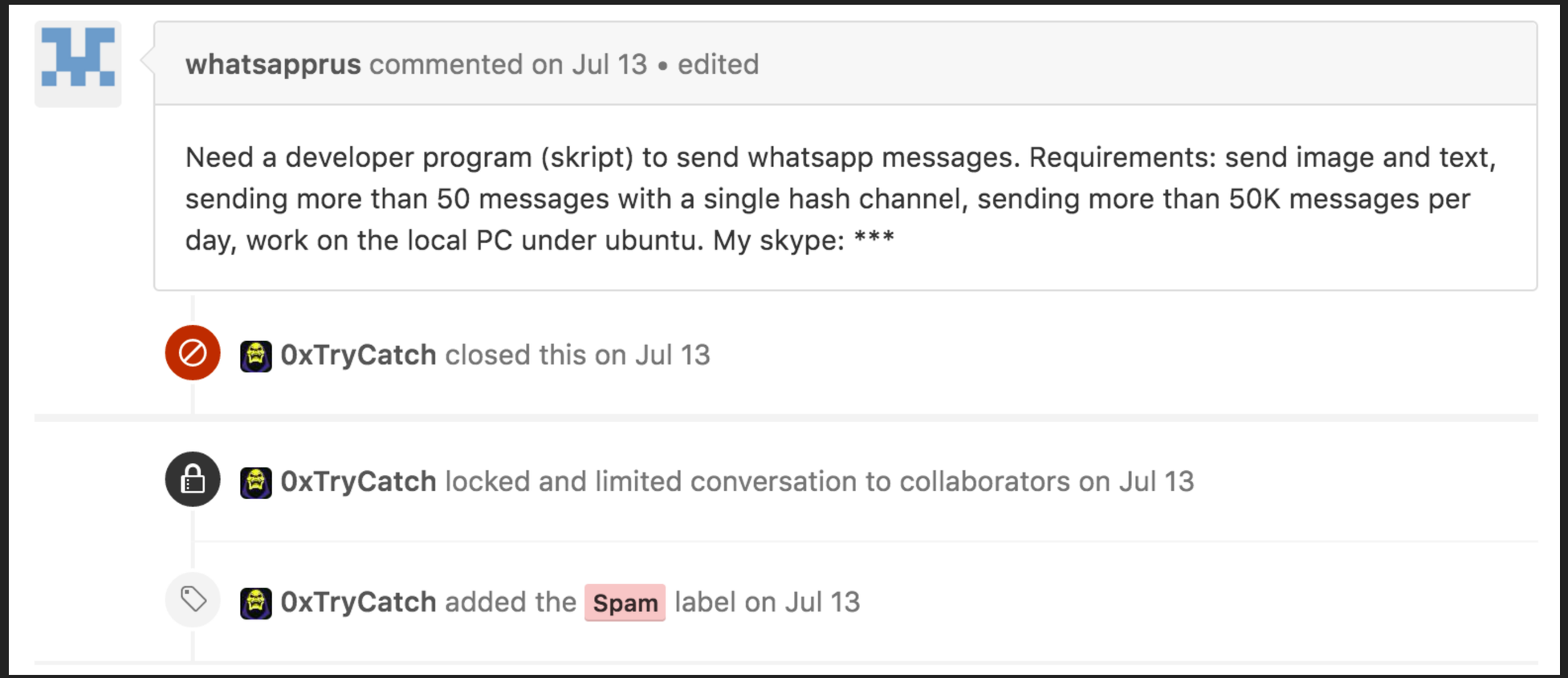
Chat API

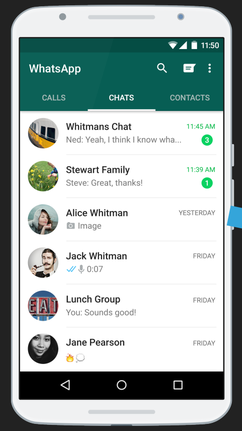
stable v3.2.0.1 | downloads 17.68 k | license GPL-3.0+

Interface to WhatsApp Messenger

Read the wiki and previous issues before opening a new one! Maybe your issue is already answered.

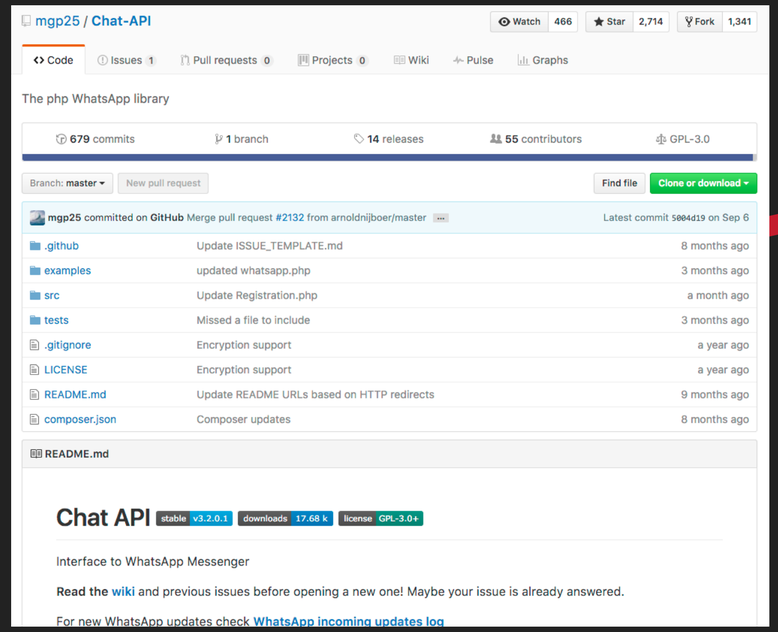
For new WhatsApp updates check **WhatsApp incoming updates log**


whatsapprus commented on Jul 13 • edited
Need a developer program (skript) to send whatsapp messages. Requirements: send image and text, sending more than 50 messages with a single hash channel, sending more than 50K messages per day, work on the local PC under ubuntu. My skype: ***
0xTryCatch closed this on Jul 13
0xTryCatch locked and limited conversation to collaborators on Jul 13
0xTryCatch added the Spam label on Jul 13


11:50
WhatsApp
CALLS
CHATS
CONTACTS
Whitmans Chat
11:45 AM
Ned: Yeah, I think I know wha...
Stewart Family
11:39 AM
Steve: Great, thanks!
Alice Whitman
YESTERDAY
Image
Jack Whitman
FRIDAY
0:07
Lunch Group
FRIDAY
You: Sounds good!
Jane Pearson
FRIDAY

?

Server


mgp25 / Chat-API
Watch 466
Star 2,714
Fork 1,341
Code
Issues 1
Pull requests 0
Projects 0
Wiki
Pulse
Graphs
The php WhatsApp library
679 commits
1 branch
14 releases
55 contributors
GPL-3.0
Branch: master
New pull request
Find file
Clone or download
mgp25 committed on GitHub Merge pull request #2132 from arnoldnijboer/master
Latest commit 5004d19 on Sep 6
.github Update ISSUE_TEMPLATE.md 8 months ago
examples updated whatsapp.php 3 months ago
src Update Registration.php a month ago
tests Missed a file to include 3 months ago
.gitignore Encryption support a year ago
LICENSE Encryption support a year ago
README.md Update README URLs based on HTTP redirects 9 months ago
composer.json Composer updates 8 months ago
README.md
Chat API
stable v3.2.0.1
downloads 17.68 k
license GPL-3.0+
Interface to WhatsApp Messenger
Read the wiki and previous issues before opening a new one! Maybe your issue is already answered.
For new WhatsApp updates check WhatsApp incoming updates log

SPAM VENDORS STOPPED



ashishnetworks commented on Apr 28

Now it has become severe issue. No use of anything. Got frustrated ! Nothing working. Getting banned every single time. We must do something about it and give this issue higher priority, otherwise there is no use of any feature if we keep getting banned.

adelsjunior commented on Aug 8

I can receive and send message , and EVEN synchronizing numbers , no day Next 'm blocked, someone KNOWS Why ?

TRAINING DATA

Features	Label
user age=5280 days, # messages 30sec=0…	Ham
user age=2 minutes, # messages 30sec=10…	Spam
user age=3 months, # messages 30sec=2…	Ham
…	…

TRAINING DATA

Features	Label
???	Ham
???	Spam
???	Ham
...	...

BEHAVIORAL FEATURES

- How do spammers generate actions?
- How do others respond to those actions?
- Classify this behavior - no access to content is fine

FEATURES: REPUTATIONAL

- "How many people have been seen on your ASN?
  What % are banned?"

FEATURES: REPUTATIONAL

- "How many people have been seen on your ASN?
  What % are banned?"

ASN:
20 total
12 banned



FEATURES: REPUTATIONAL

- "How many people have been seen on your Netblock?
  What % are banned?"

ASN:
20 total
12 banned



Netblock:
8 total
5 banned



FEATURES: REPUTATIONAL

- "How many people have been seen on your Phone Prefix? What % are banned?"

ASN:
20 total
12 banned



Netblock:
8 total
5 banned



Phone Prefix:
30 total
23 banned



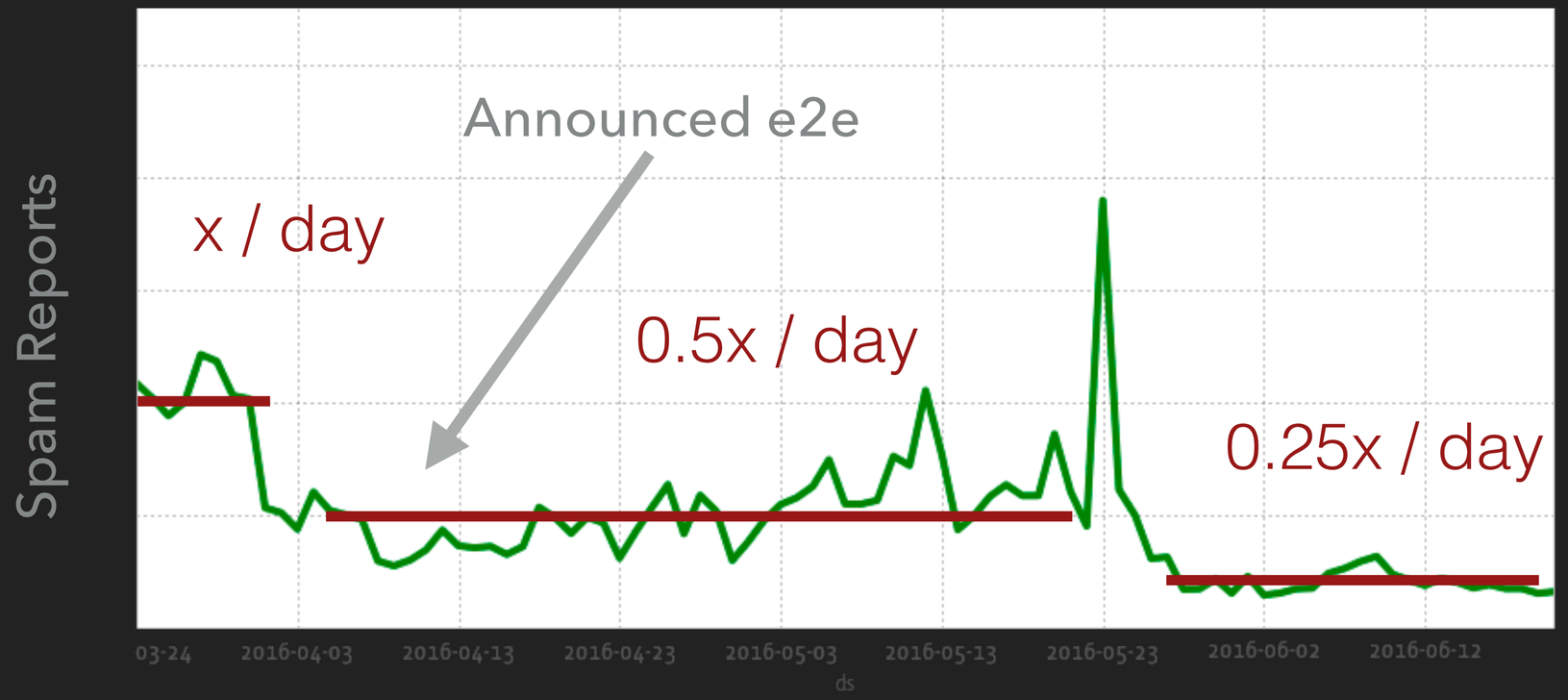
FEATURES: REPUTATIONAL

- "How many people have been seen on your ASN?
  What % are banned?"
- Generalize “ASN”: “netblock”, “phone number prefix”, etc
- Generalize “banned”: “reported as spam”, etc
- Forces attacker to buy more things

FEATURES: PROBABILITY

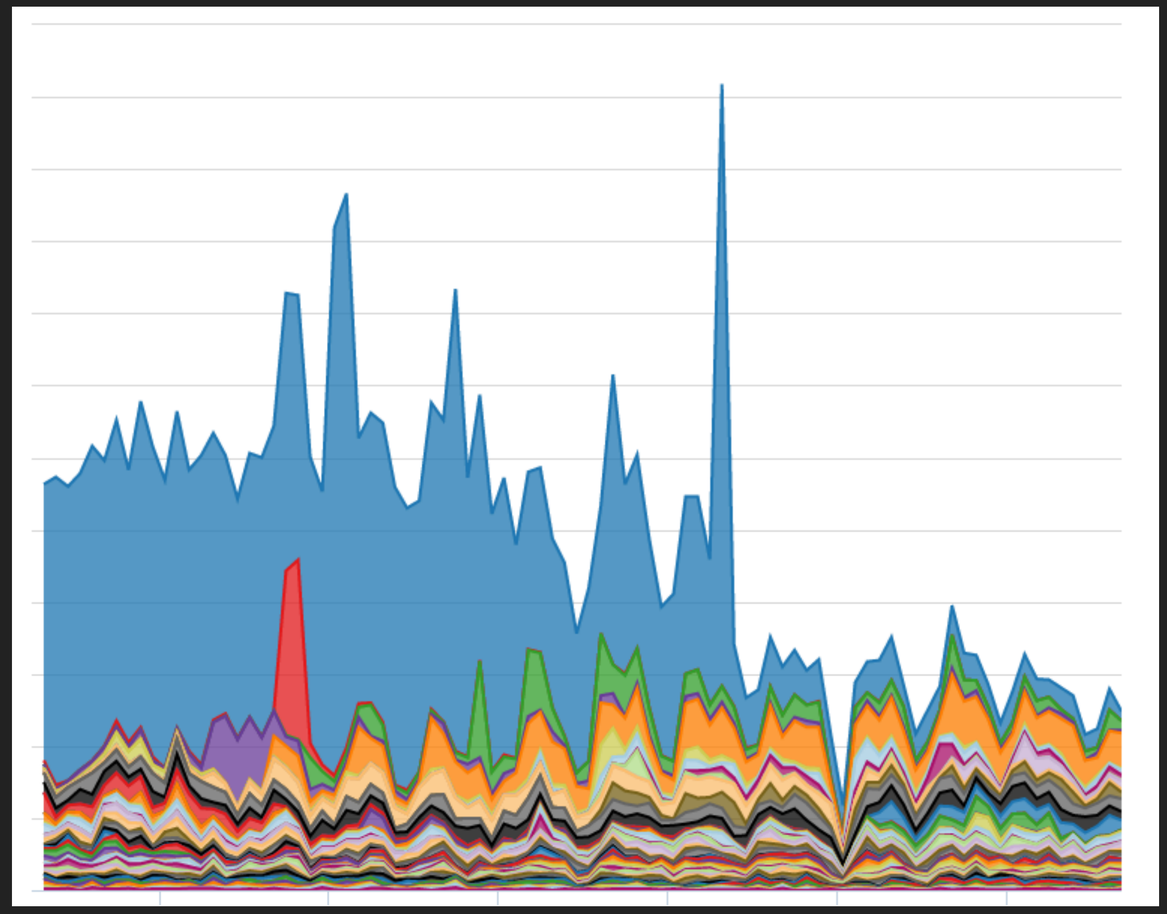
- “Probability that someone with phone country code X is seen from network country Y”
- Generalize each (carrier, language, etc)

SPAM REPORTS DECREASED



BAN APPEALS DECREASED

Appeals



Time (3 months)

YOU DON'T NEED CONTENT FOR SPAM DETECTION

- 2 keys to success:
  - Good behavioral features
  - Good labeling: Spam Reports + Spam App detection