Tools for WhatsApp Data Collection

Kiran Garimella (MIT), Philipe Melo (UFMG), Gareth Tyson (QMUL), Jussara Almeida, Fabricio Benevenuto (UFMG)

http://tiny.cc/ICWSMWhatsApp

#WhatsAppICWSM

Housekeeping

Disclaimer

- Collecting WhatsApp data is not easy!
- This tutorial is not complete!
- We'll try to keep it non technical
- But sometimes we'll have details
  - Engineering details/knowledge required

Disclaimer

- This is for **public** WhatsApp groups

- One-to-one messaging...
  - NO
- Private group messaging...
  - NO
- Example public group: http://tiny.cc/icwsmgroup

Disclaimer

- This is all according to the TOS (apart from the copyright logo infringement)
- Important to keep ethical considerations in mind (We have IRB approval)
- Preliminary analysis

Feel free to download

Code: https://github.com/Phlop/WhatsApp_Tutorial/

Colab: https://colab.research.google.com/drive/18-lcIaDdOQ3SEaUaZy-J08rVFMqtpHkj (create a copy and play with the scripts)

Toolkit Paper: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17865/17043 (WhatsApp Doc?A First Look at WhatsApp Public Group Data from ICWSM'18)

Agenda

1. What's WhatsApp? (15 mins)
2. What data can we get (30 mins)
3. How do we get it (30 mins)

Break

---

4. What can we do with it (1 hour)
   - Case studies
5. Discussion (30 mins)

Part 1: What’s WhatsApp?

What's WhatsApp?

- A messaging platform for the web & mobile apps
- Similar to SMS on steroids
- Integrated multimedia services

WhatsApp 101

- Mainly a 1-1 messaging platform
- But ~10% of the messages are 1-many (groups)
- Multimedia heavy
- End-to-end encrypted

This is really cool to study...

Why WhatsApp?

- Different platform & use cases
- Different content (multimedia & text)
- Different demographics
- Has an effect on education, social services, governance, social fabric

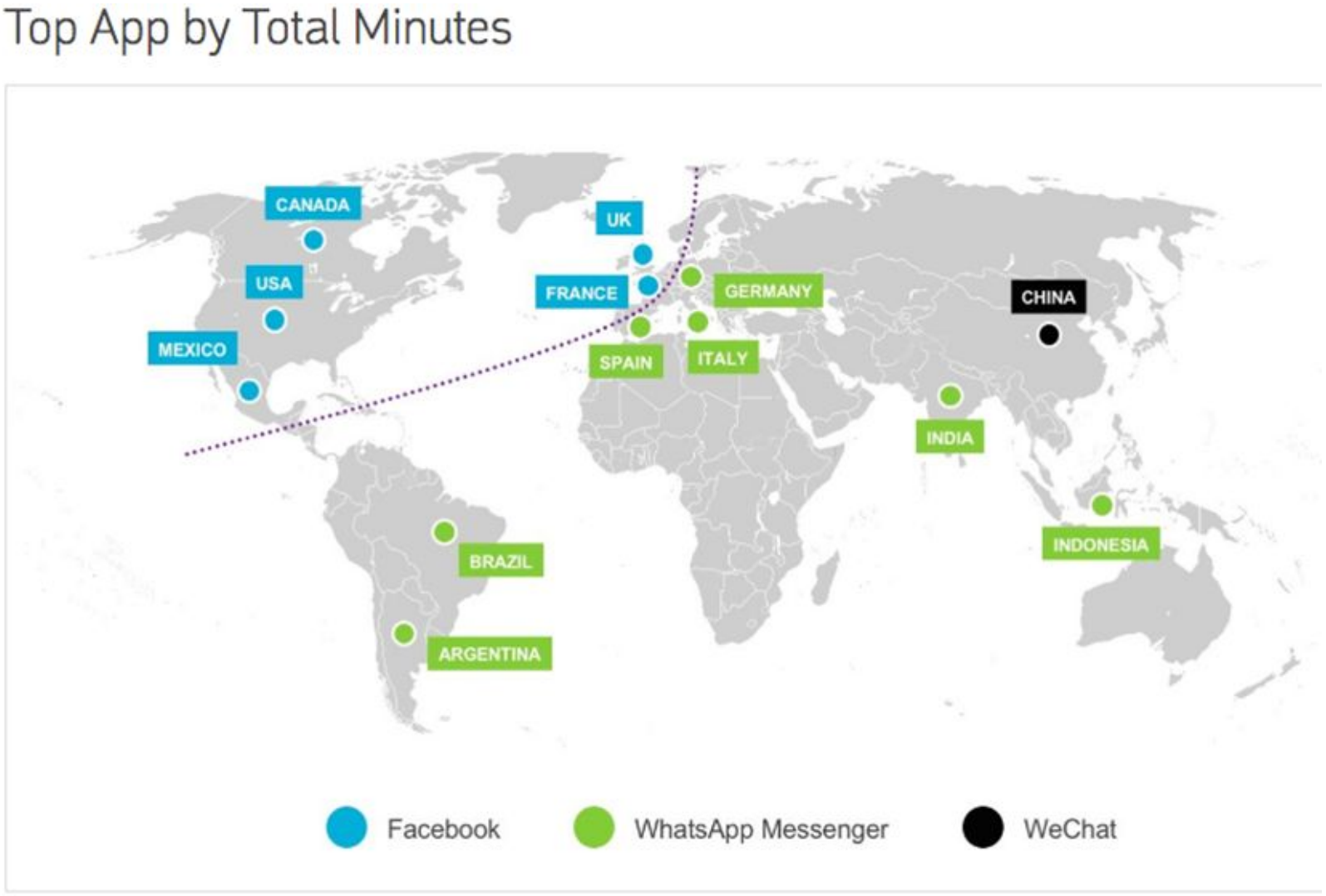
WhatsApp is BIG

- Over 1.5 Billion monthly active users
- 500 million daily active users

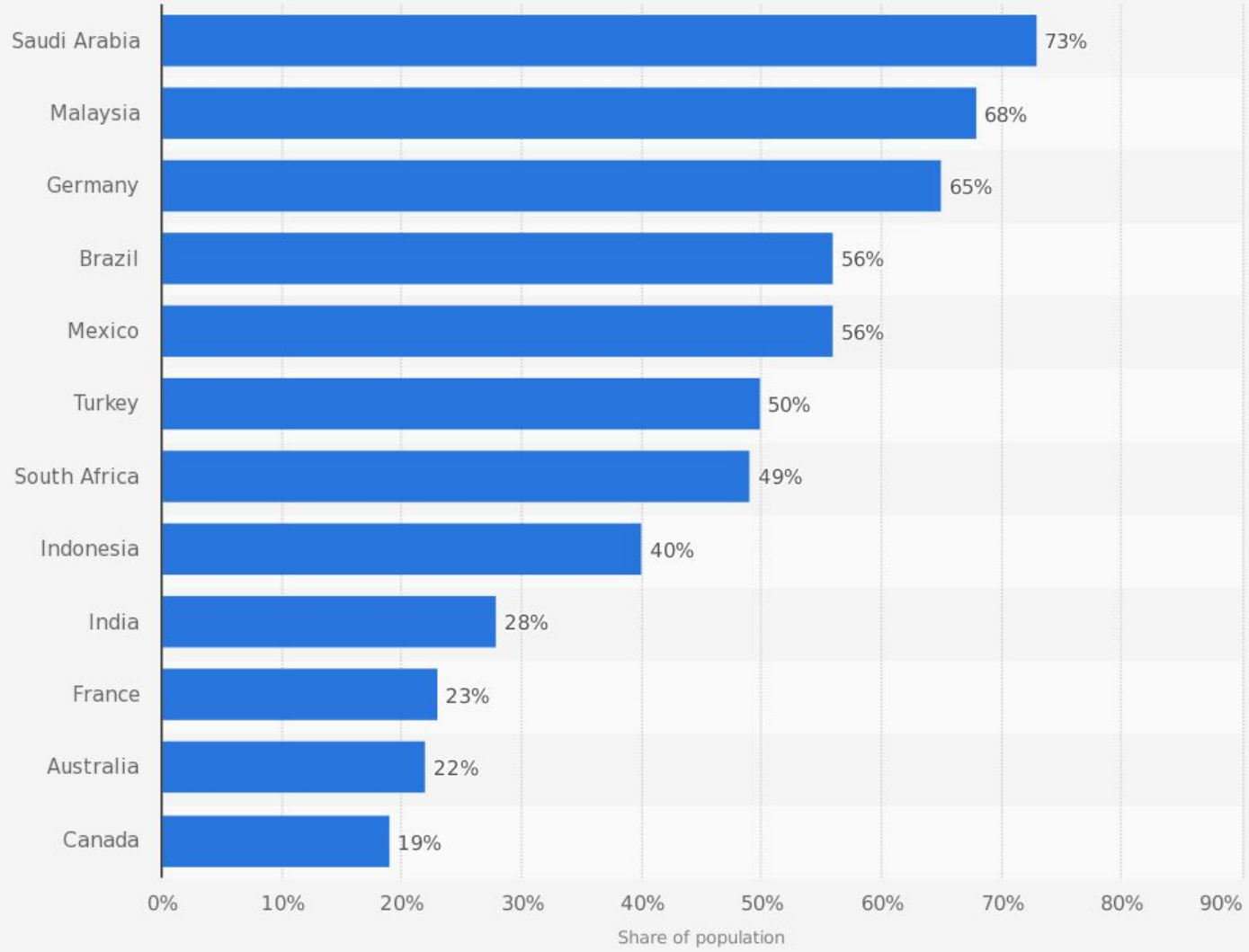
- Over 60 billion texts, 100 million audio and 55 million video calls daily

- 4 billion minutes of communication calls
- 5 billion pictures shared daily

WhatsApp Stats



Share of population in selected countries who are active WhatsApp users as of 3rd quarter 2017



Sources
GlobalWebIndex; We Are Social
© Statista 2019

Additional Information:
Worldwide; GlobalWebIndex; We Are Social; Q2 2017 to Q3 2017; 16 to 64 years

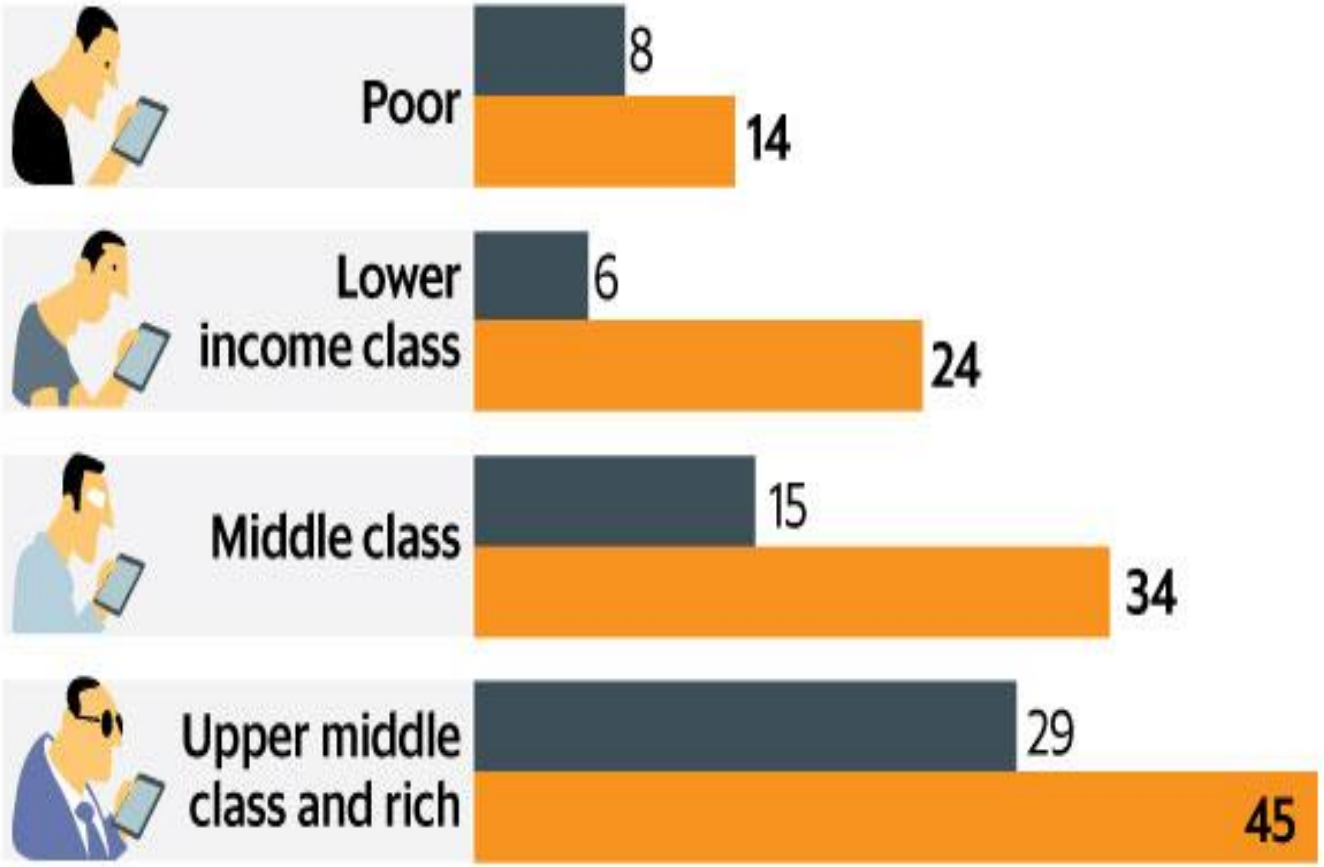
WhatsApps in Emerging Markets

- High penetration!
  - 95% of Android comes with WhatsApp pre-installed
- Replacement for the internet in Brazil, Mexico, Indonesia, India
- Over 300 million Indians, 120M Brazilians on WhatsApp
- In Zimbabwe, WhatsApp was responsible for half of all internet traffic in the country last year.

Chart 2

The reach of WhatsApp extends beyond the rich

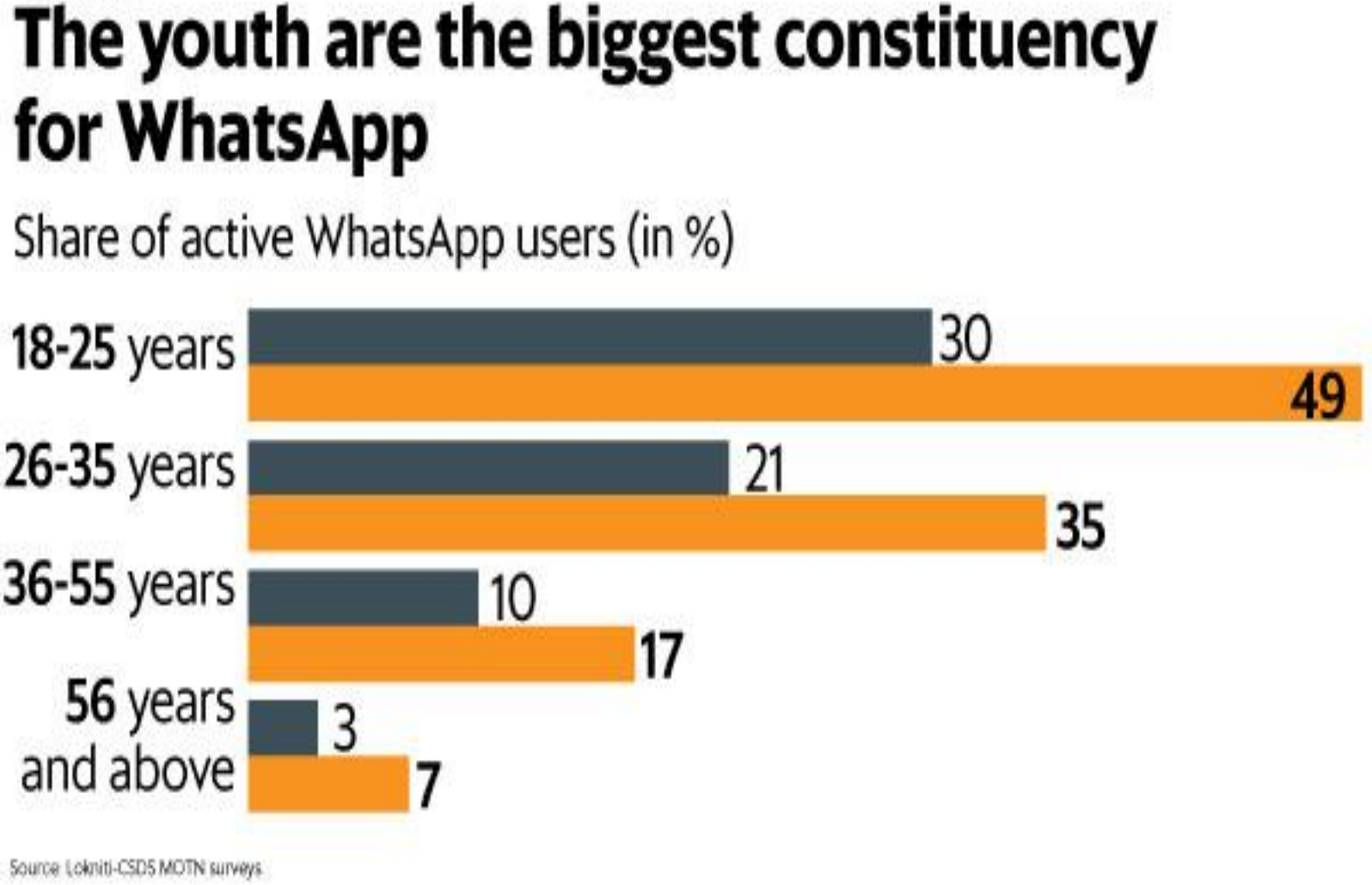
Share of active WhatsApp users (in %)



Source: Lokniti-CSDS MOTN surveys

Chart 3

The youth are the biggest constituency for WhatsApp

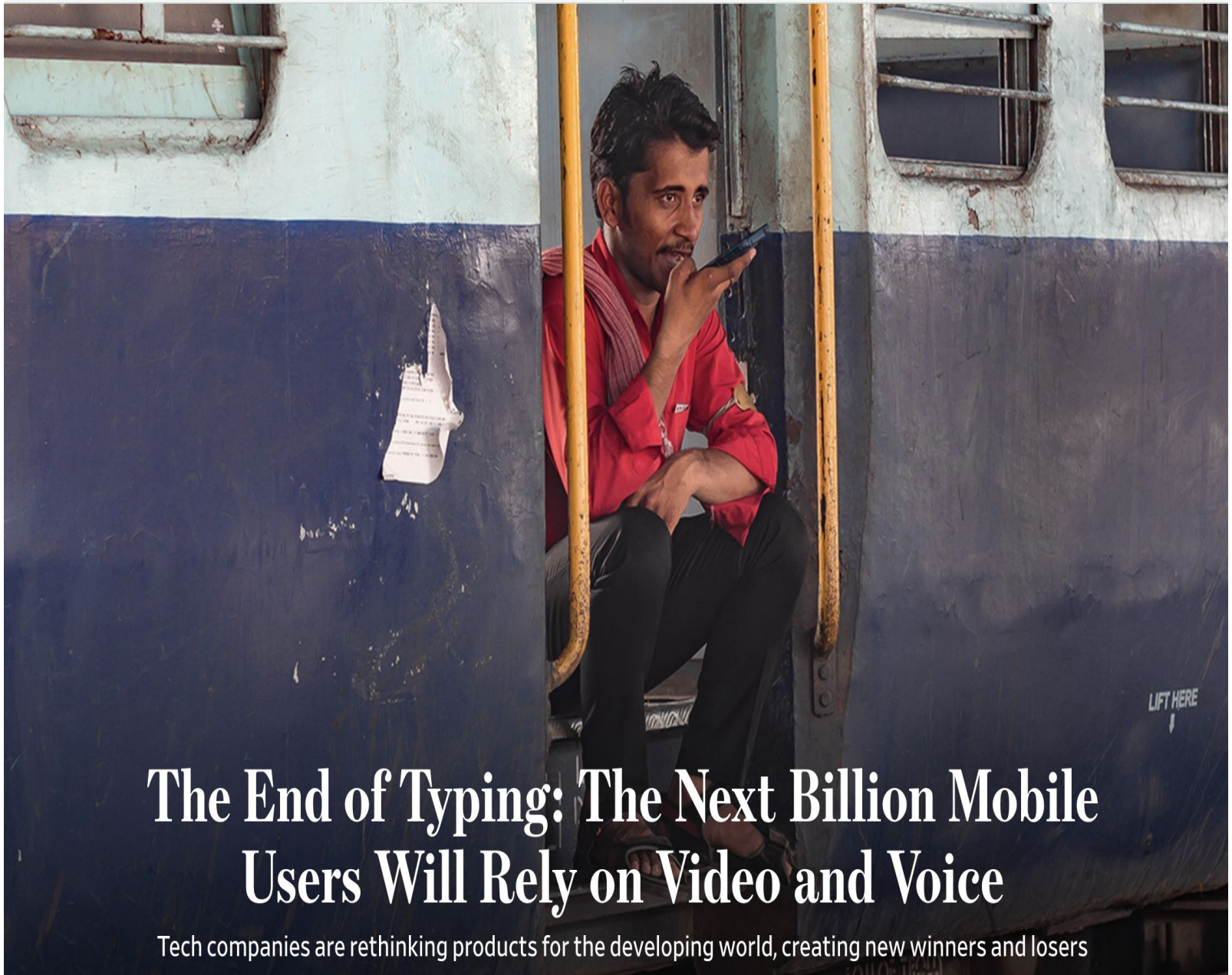


Source: Lokniti-CSDS MOTN surveys

This is starting to get wide media attention

THE WALL STREET JOURNAL.

SUBSCRIBE NOW SIGN IN



The End of Typing: The Next Billion Mobile Users Will Rely on Video and Voice

Tech companies are rethinking products for the developing world, creating new winners and losers

Support The Guardian
Contribute →
Search jobs | Sign in | Search | International edition
The Guardian
News | Opinion | Sport | Culture | Lifestyle | More
World ▸ Europe US Americas Asia Australia Middle East Africa Inequality Cities Global development

Guardian Africa network
Sudan

How WhatsApp is fuelling a 'sharing revolution' in Sudan

Khalid Albaih

Thu 15 Oct 2015 09.00 BST

Amid a government crackdown on the press, the app is helping politically charged residents share news and views without fear

Stoxx 600 364.41 -0.36% ▼ Nikkei 22269.88 0.09% ▲ U.S. 10 Yr 0/32 Yield 3.184% ▼ Crude Oil 60.75 0.93% ▲ Euro 1.1255 -0.71% ▼ DJIA 25989.30 -0.77% ▼

THE WALL STREET JOURNAL.

A-HED

The Internet Is Filling Up Because Indians Are Sending Millions of 'Good Morning!' Texts

The country, newly online, has become obsessed with sending cheery greetings from smartphones, causing head scratching in Silicon Valley

By *Newley Purnell*

Jan. 22, 2018 11:54 a.m. ET

Google researchers in Silicon Valley were trying to figure out why so many smartphones were freezing up half a world away. One in three smartphone users in India run out of space on their phones daily.

Most Popular Videos

1. What Trump Did to Silence Stormy Daniels and Karen McDougal

Smartphones, Facebook, WhatsApp break down marriages in Bangalore

WORLD • INDIA

Five Killed in Latest Mob Attack After Rumors on Social Media. Here's What to Know About India's WhatsApp Murders

Sign Up for Our Newsletters

Sign up to receive the top stories you need to know now on politics, health, money and more

WORLD

WhatsApp Has Become A Hotbed For Spreading Nazi Propaganda In Germany

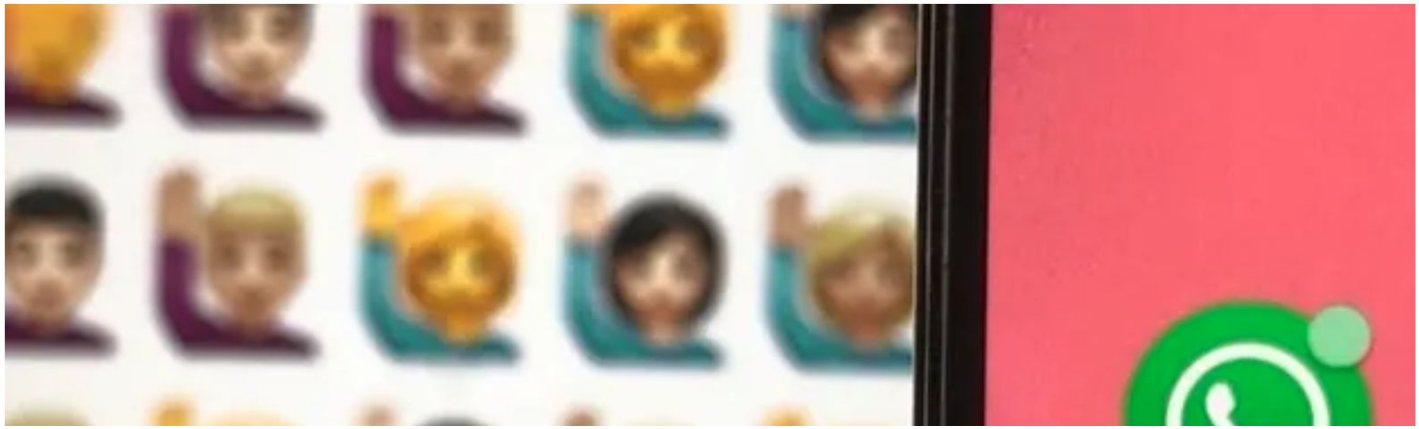
BuzzFeed News has documented hundreds of uses of anti-Semitic or otherwise pro-Nazi stickers used in WhatsApp groups, despite those images being illegal in Germany.

Karsten Schmehl
BuzzFeed-Redaktion, Deutschland

Reporting From
Berlin, Germany

Posted on April 18, 2019, at 7:13 a.m. ET

Tweet Share Copy



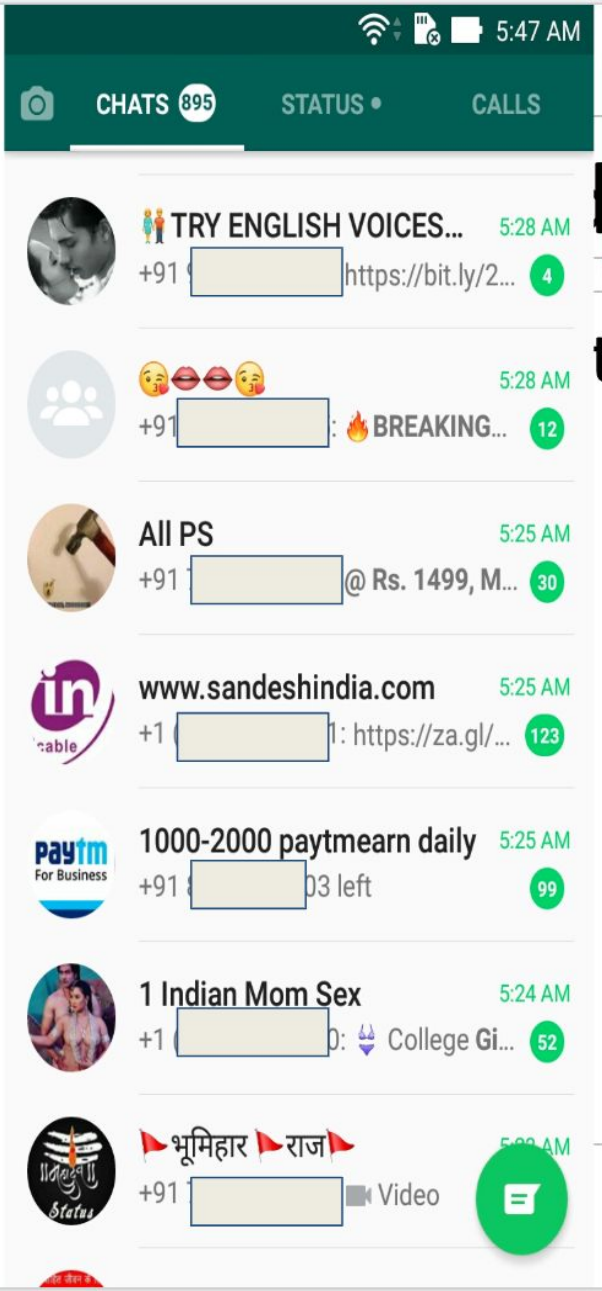
Part 2: What data can we get?

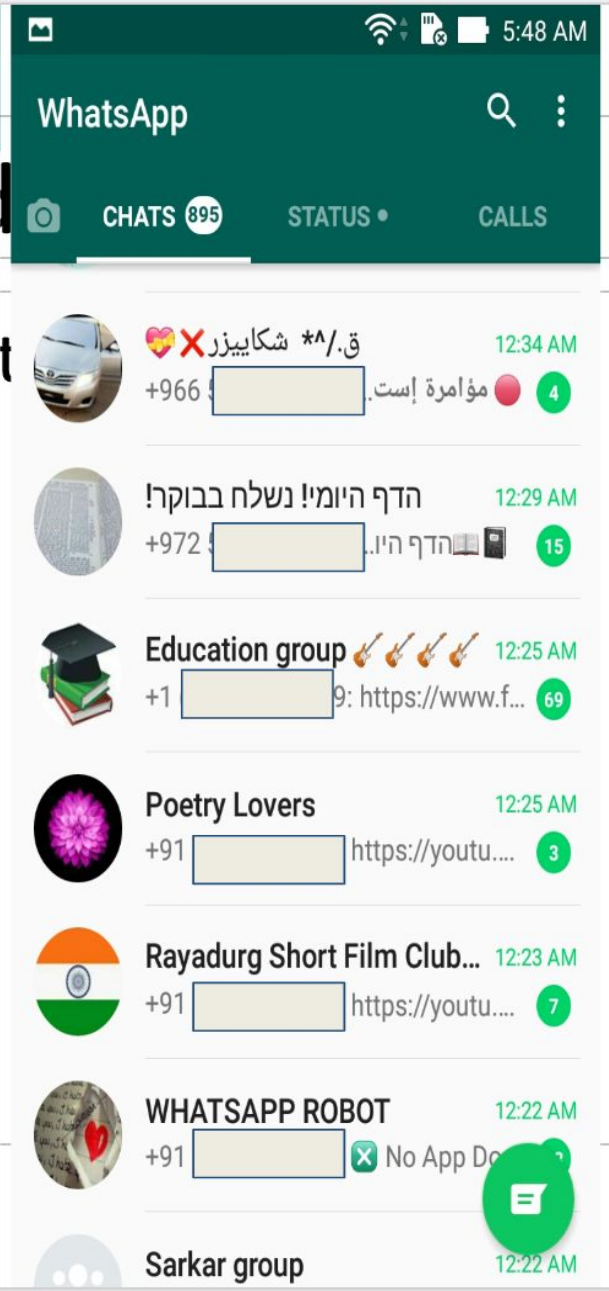
WhatsApp Public groups

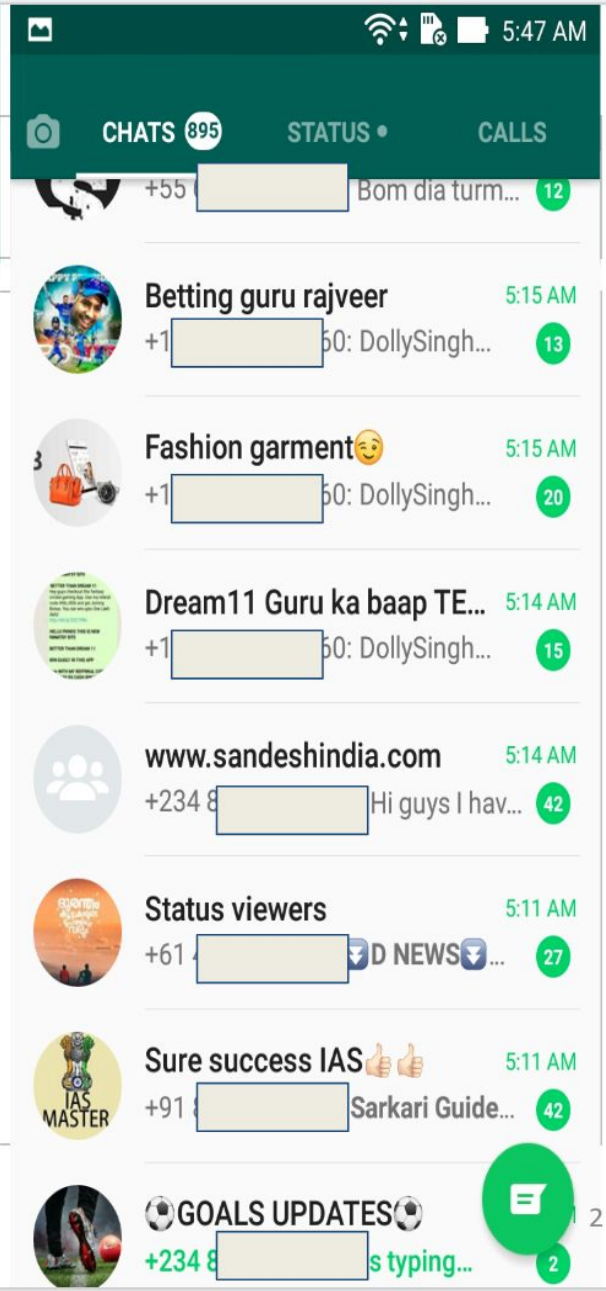
- Public groups
  - Around 10-20% of the communication is through groups.
  - Some of them are public groups.
  - Anyone with a group link (chat.whatsapp.com/xyz) can join.
  - Limited to 256 members per group.
- Public lists of groups
  - Public websites e.g. https://www.whatsappgrouplinks.org/
  - Google search
  - Facebook/Twitter

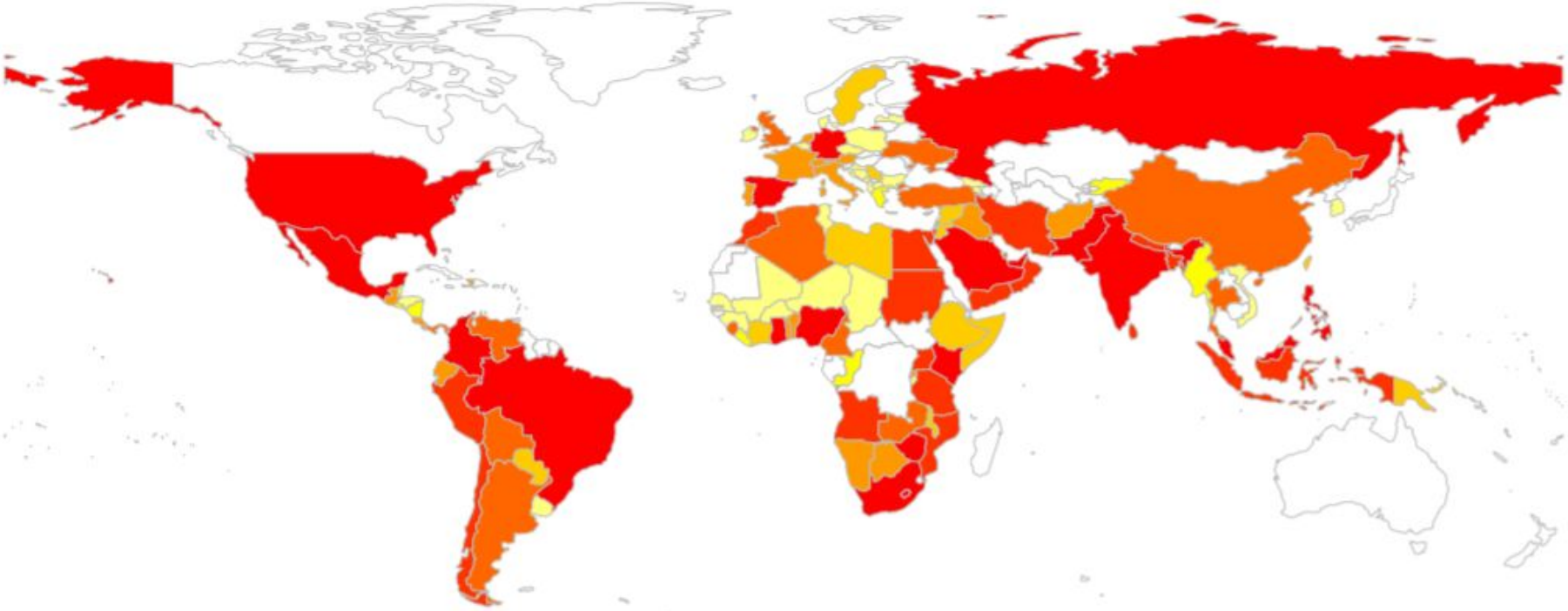
WhatsApp data collection

- Collected 100k public groups
- Requires manual work
  - Most of them are not useful.
  - Needs to be targeted and specific.
- Joined a random set of 1,500 of them
- Over a million messages from ~100k users collected in 3 months.


5:47 AM
CHATS 895
STATUS •
CALLS
TRY ENGLISH VOICES... 5:28 AM
+91 https://bit.ly/2... 4
5:28 AM
+91 : BREAKING... 12
All PS 5:25 AM
+91 @ Rs. 1499, M... 30
www.sandeshindia.com 5:25 AM
+1 1: https://za.gl/... 123
1000-2000 paytmearn daily 5:25 AM
+91 03 left 99
1 Indian Mom Sex 5:24 AM
+1 0: College Gi... 52
भूमिहार राज
+91 Video


5:48 AM
WhatsApp
CHATS 895
STATUS •
CALLS
ق./*٨ شكاييزر 12:34 AM
+966 مؤامرة إست. 4
הדף היומי! נשלח בבוקר! 12:29 AM
+972 הדף היו... 15
Education group 12:25 AM
+1 9: https://www.f... 69
Poetry Lovers 12:25 AM
+91 https://youtu.... 3
Rayadurg Short Film Club... 12:23 AM
+91 https://youtu.... 7
WHATSAPP ROBOT 12:22 AM
+91 No App Do...
Sarkar group 12:22 AM


5:47 AM
CHATS 895
STATUS •
CALLS
+55 Bom dia turm... 12
Betting guru rajveer 5:15 AM
+1 60: DollySingh... 13
Fashion garment 5:15 AM
+1 60: DollySingh... 20
Dream11 Guru ka baap TE... 5:14 AM
+1 60: DollySingh... 15
www.sandeshindia.com 5:14 AM
+234 8 Hi guys I hav... 42
Status viewers 5:11 AM
+61 D NEWS... 27
Sure success IAS 5:11 AM
+91 Sarkari Guide... 42
GOALS UPDATES
+234 8 s typing... 2

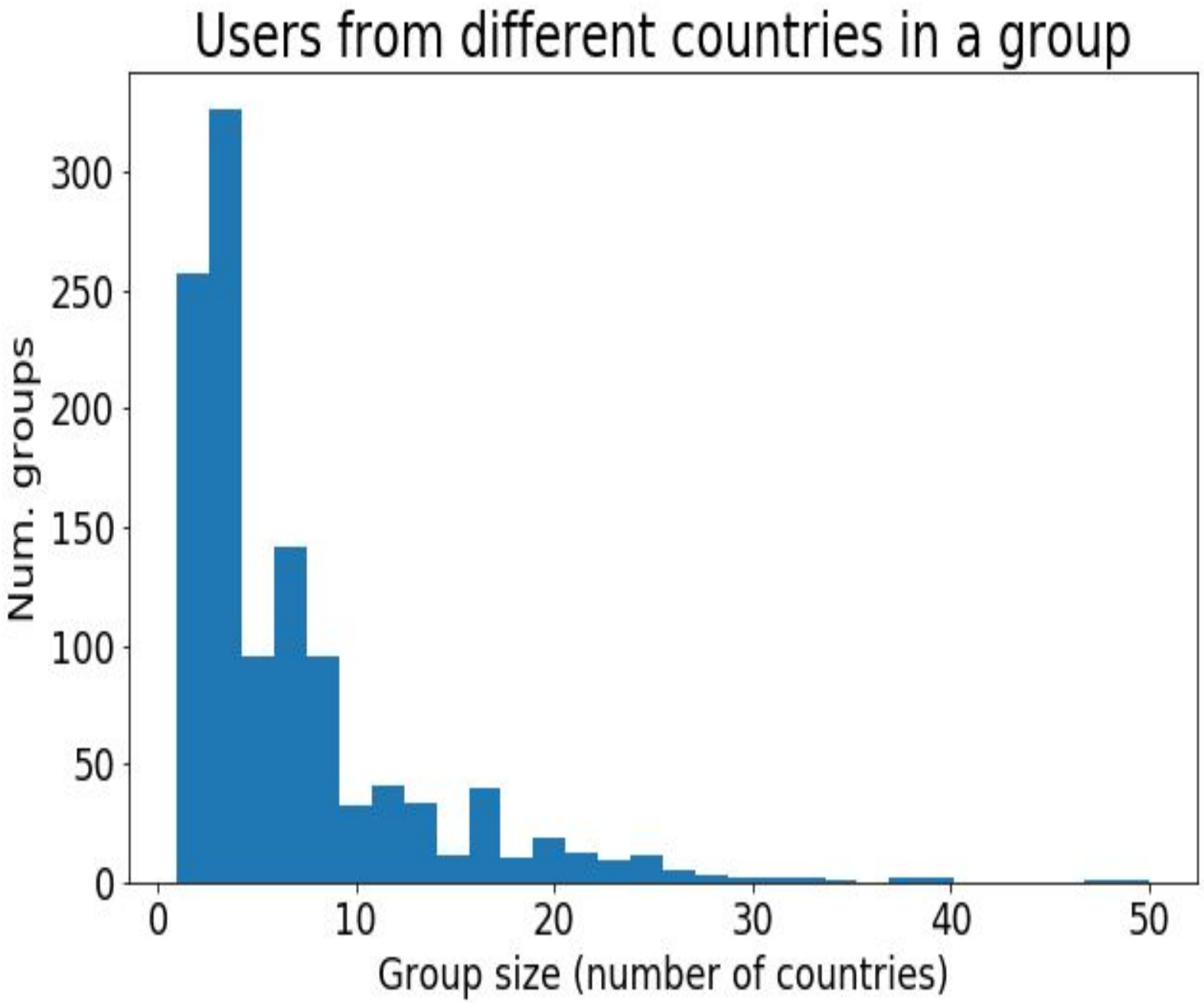


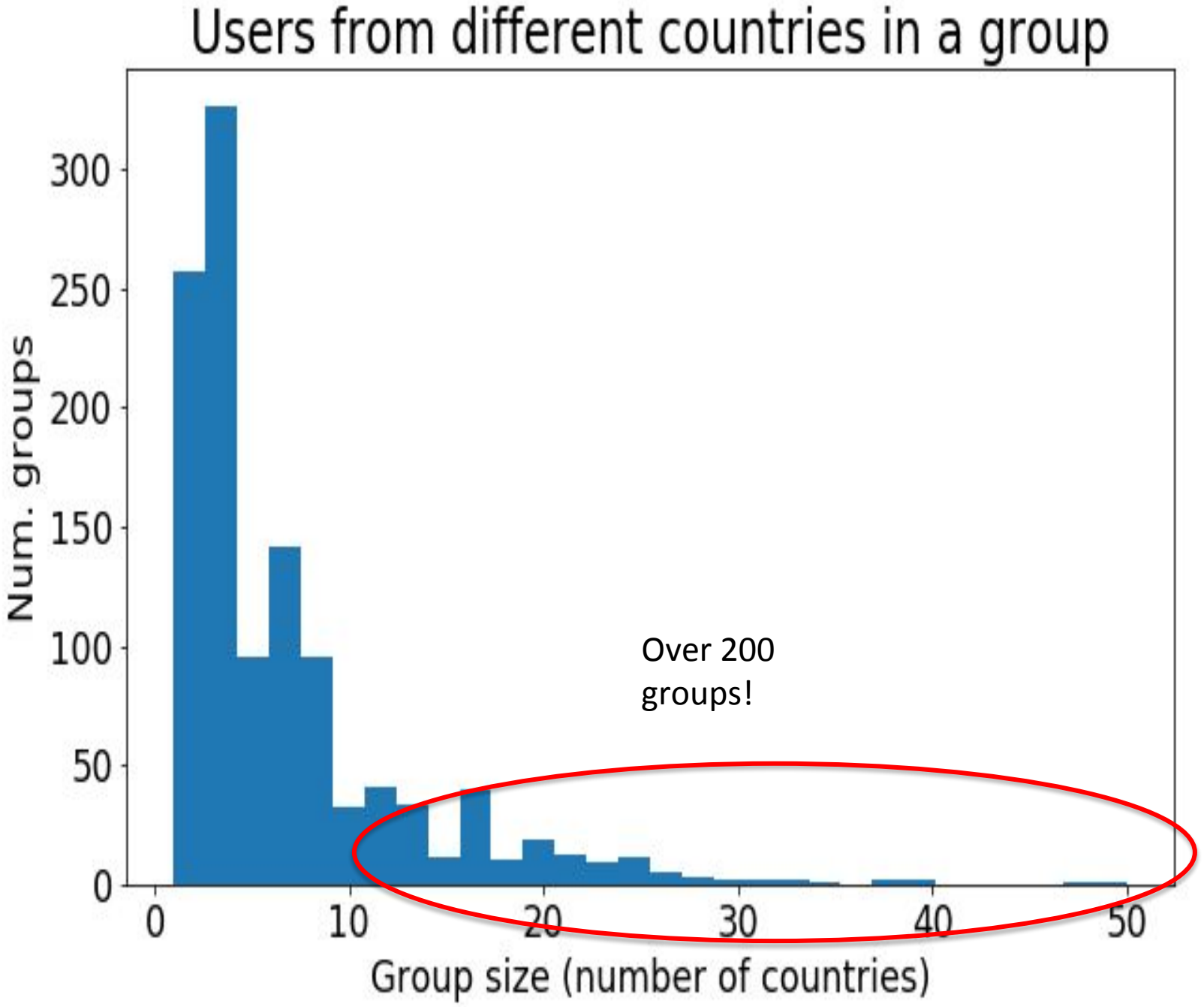
What's in the data?

- All messages in the group

- Images, video, audio and PDF documents
- Names, phone numbers

- Business accounts
  - Email, address
  - E.g. professional services, restaurants, health, grocery, event planning, government, etc


Users from different countries in a group
Num. groups
300
250
200
150
100
50
0
0
10
20
30
40
50
Group size (number of countries)


Users from different countries in a group
Num. groups
300
250
200
150
100
50
0
0
10
20
30
40
50
Group size (number of countries)
Over 200 groups!

Now over to…..

…Kiran!

Part 3: Data collection

Join our group: http://tiny.cc/icwsmgroup

to be part of the data demonstration.

WhatsApp data collection

- Need an Android phone
- Details here: https://github.com/gvrkiran/whatsapp-public-groups

Requirements

1. Functional Whatsapp account
2. Working smartphone
3. Virtual machine (optional)

Data collection in this tutorial

Public Group based

Most of the techniques generalize for any groups

For many other mobile apps (e.g. telegram, kik, discord, etc)

Data collection steps

1. Find relevant groups
2. Join them on the phone
3. Backup the database
4. Extract messages
    a. Download images, videos, etc.

Data collection steps

1. Find relevant groups <- semi-manual
2. Join them on the phone <- automated
3. Backup the database <- semi automated
4. Extract messages <- automated
    a. Download images, videos, etc.

Caution!

TMI!!

Too **M**any eng**I**neering details. ;)

You can skip to section 4 if you are not interested in the engineering details

Code here

https://tinyurl.com/whatsapp-tutorial

Finding the groups

- Follow steps in Task 1 in the google collab.
  - Get list of groups from google search
    - and get their titles automatically.
  - Get list of groups manually
    - http://whopostedwhat.com (facebook search)
    - Twitter/reddit search
    - https://joinwhatsappgroup.com/
    - and many others.

Getting data

4 ways:

Manually

Web WhatsApp

Rooted phone

Jail broken WhatsApp

Getting data

4 ways:

Manually

Web WhatsApp

Rooted phone

Jail broken WhatsApp

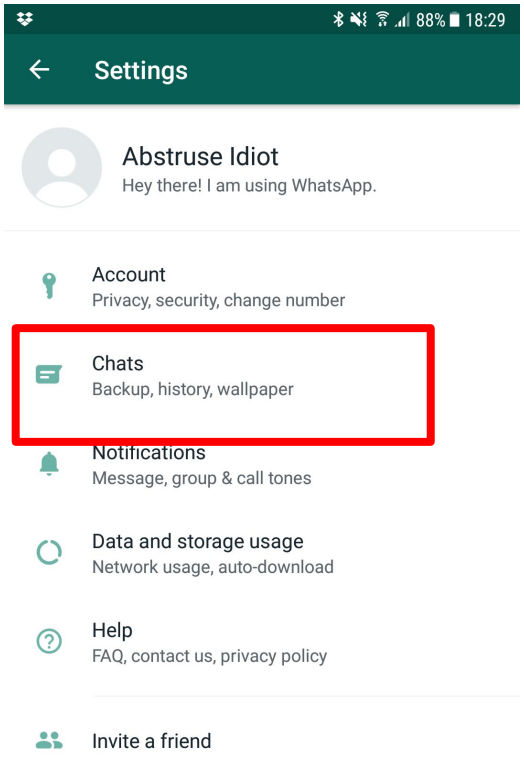
Getting data

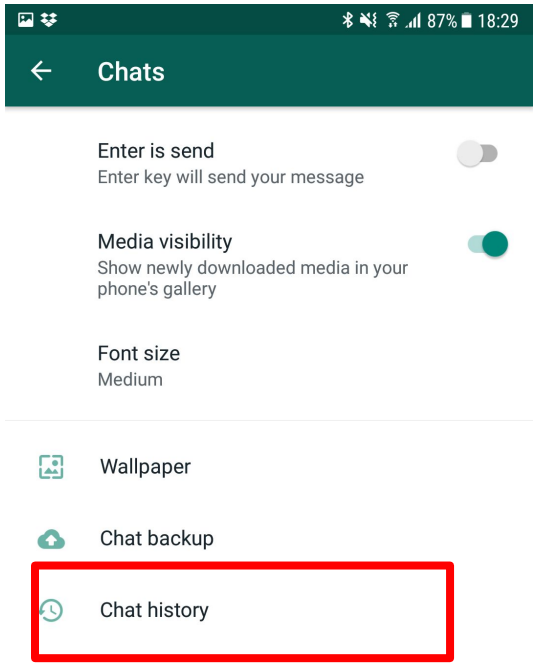
4 ways:

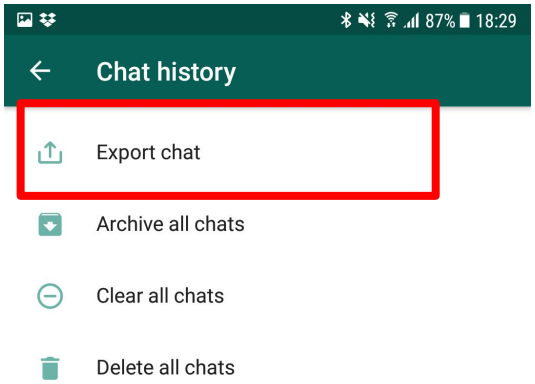
Manually	Web WhatsApp	Rooted phone	Jail broken WhatsApp
Not scalable Tedious	Fragile Breaks frequently	Unsafe	Unsafe

Manual

- Go to Settings -> Chats -> Chat history -> Export chat
- Have to do it manually for all the group chats

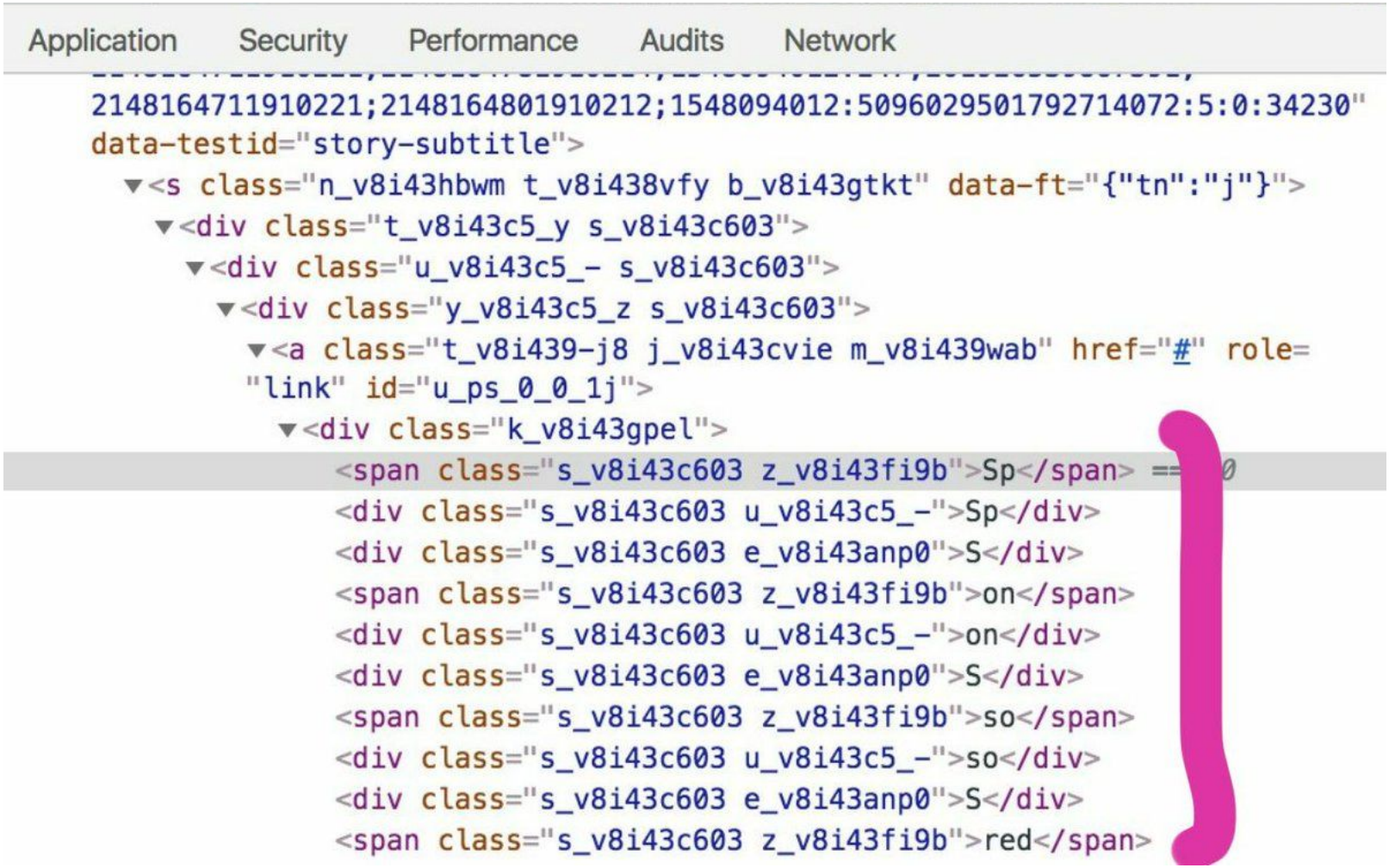






Web WhatsApp

- WhatsApp provides a web interface
- Use scraping/automation to get information
- Tools exist:
  - https://github.com/mukulhase/WebWhatsapp-Wrapper
  - Works decently, we used it
  - But extremely fragile
  - Could stop working if WhatsApp changes structure


Application
Security
Performance
Audits
Network
2148164711910221;2148164801910212;1548094012:5096029501792714072:5:0:34230"
data-testid="story-subtitle">
<s class="n_v8i43hbwm t_v8i438vfy b_v8i43gtkt" data-ft="{"tn":"j"}">

<a class="t_v8i439-j8 j_v8i43cvie m_v8i439wab" href="#" role=
"link" id="u_ps_0_0_1j">

Sp
Sp
S
on
on
S
so
so
S
red

Rooted phone

- WhatsApp stores the database containing all the data on our phone locally.
- But the database is encrypted.
- Look for your encrypted database at:
  - device storage -> WhatsApp -> Databases -> msgstore.db.crypt12
  - browse here: http://www.whatsapp-monitor.dcc.ufmg.br/Android/index.html
- We need to have root access to access the database or the key for decryption.

Rooted phone

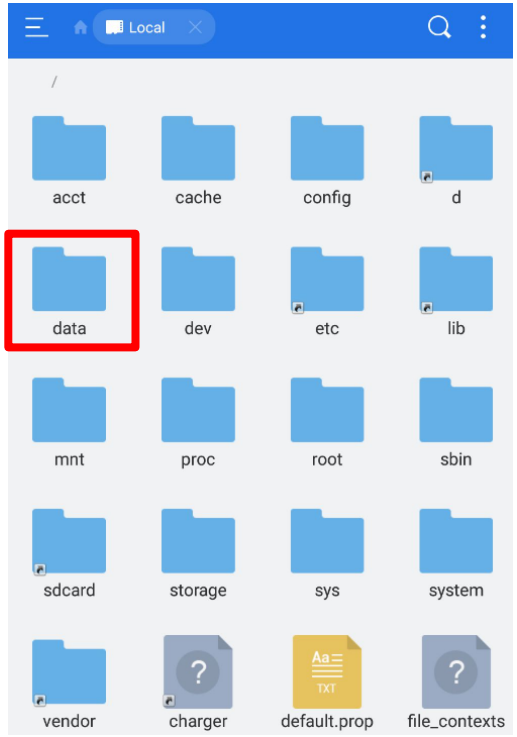
- STOP!
- Rooting a phone is not recommended. It can create security issues.
- Use a stand alone phone.
  - or a Virtual Machine.

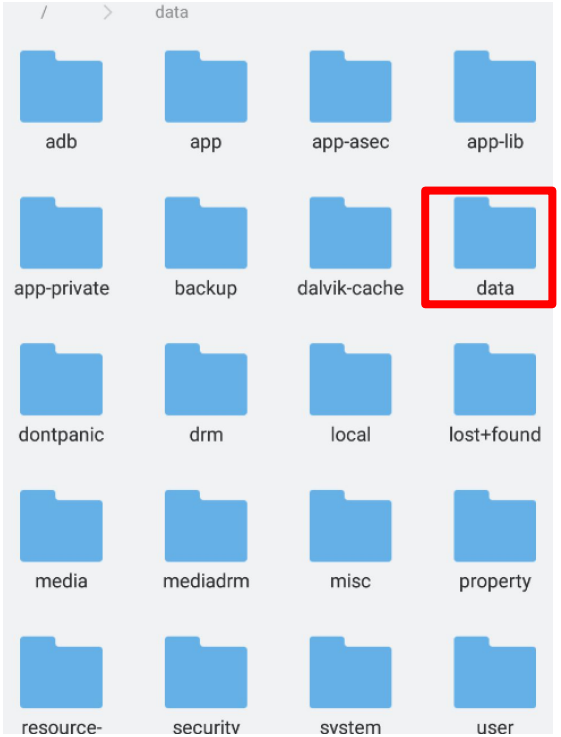
Where can I find the databases?

-

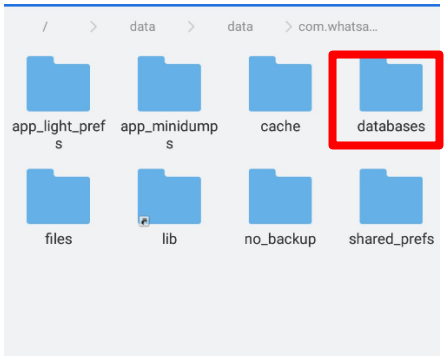
Where can I find the databases?

Row #	Content	Directory	File
1	contacts database	/data/data/ com.whatsapp/databases	wa.db (SQLite v.3)
2	chat database	/data/data/ com.whatsapp/databases	msgstore.db (SQLite v.3)
3	backups of the chat database	/mnt/sdcard/ Whatsapp/Databases	msgstore.db.crypt msgstore-<date>.crypt
4	avatars of contacts	/data/data/ com.whatsapp/files/ Avatars	UID.j, where UID is the identifier of the contact
5	copies of contacts avatars	/mnt/sdcard/ What-sApp/ProfilePictures	UID.j, where UID is the identifier of the contact
6	log files	/data/data/ com.whatsapp/files/ Logs	whatsapp.log, whatsapp-<date>.log
7	received files	/mnt/sdcard/ Whatsapp/Media	various files
8	sent files	/mnt/sdcard/ Whatsapp/Media/Sent	various files
9	user settings and preferences	/data/data/ comm.whatsapp/files	various files





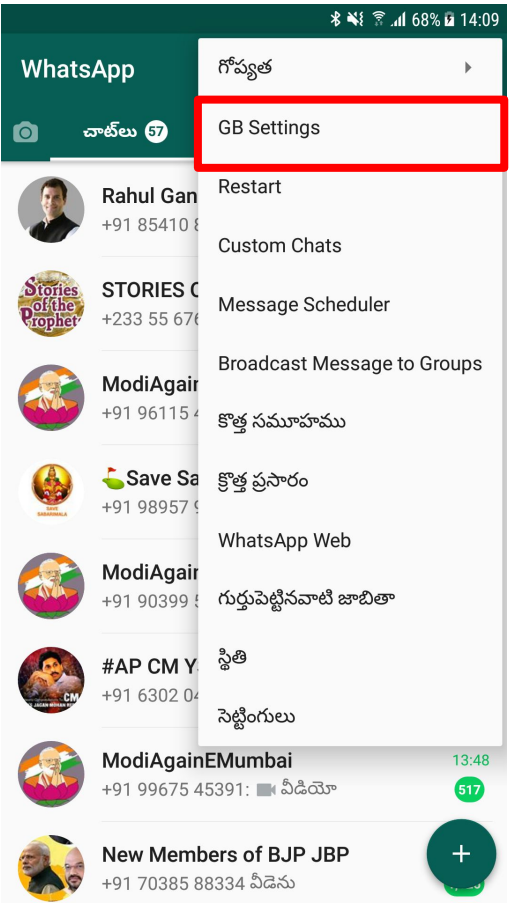
Data in:
/data/data/com.whatsapp.com/Databases/

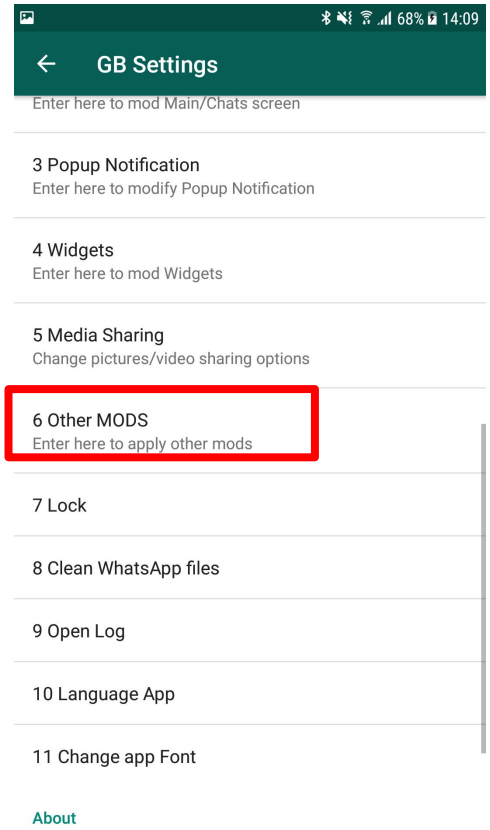


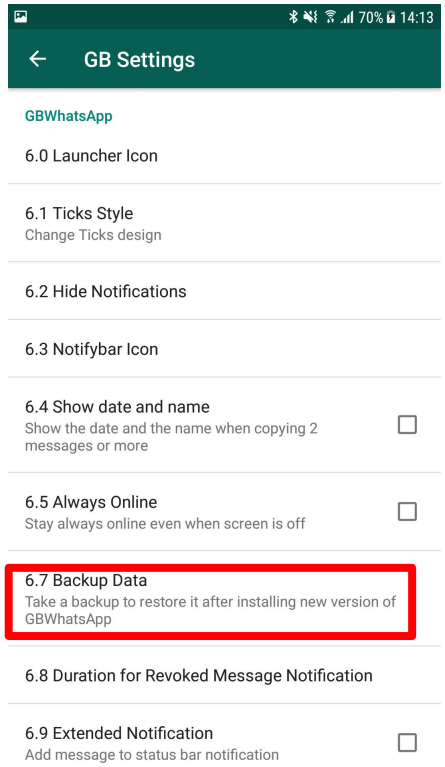
Jailbroken WhatsApp

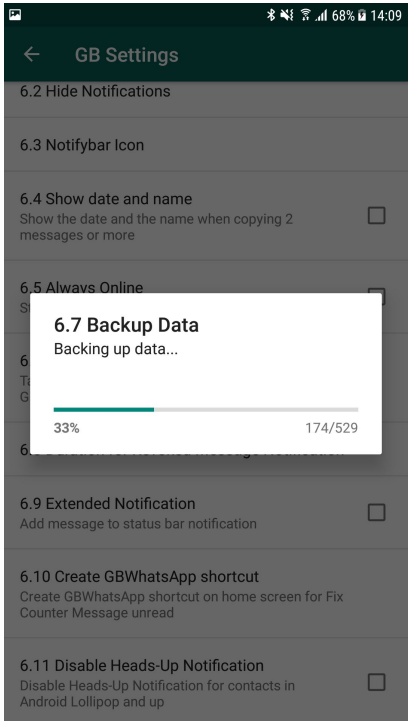
- Again, not recommended, unless you have a standalone phone
- Famous ones include GBWhatsApp, YoWhatsApp, etc.

Jail broken whatsapp

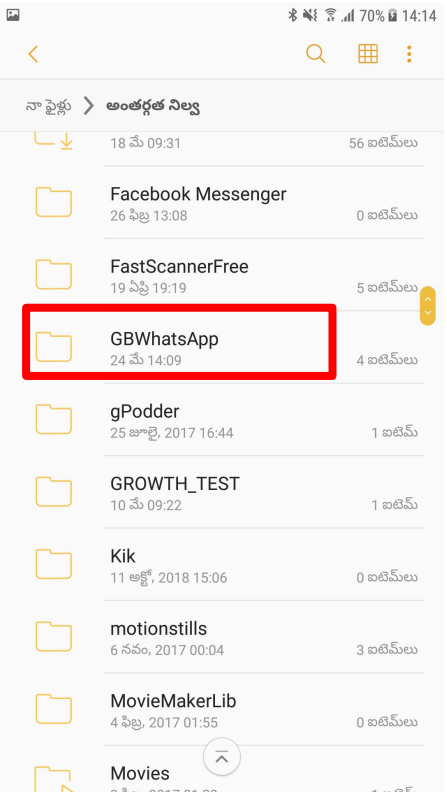


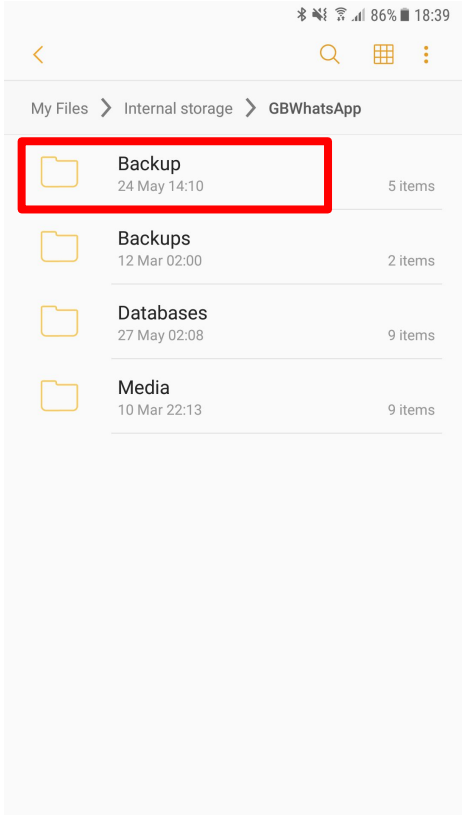


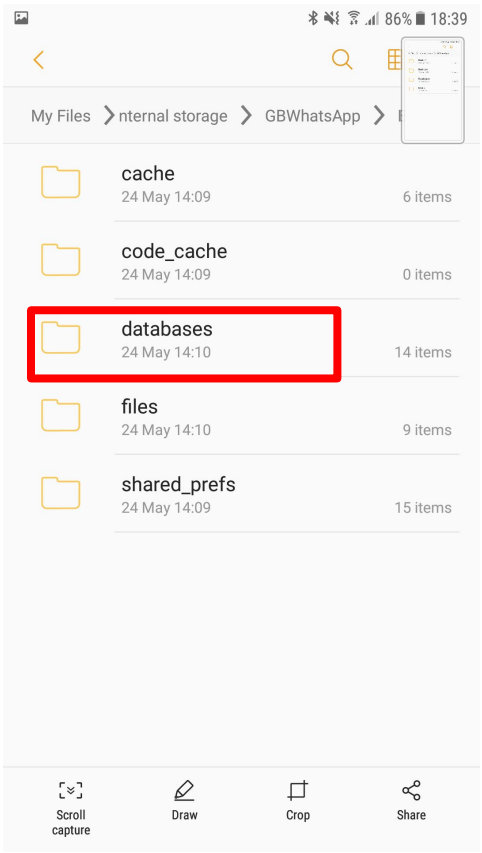


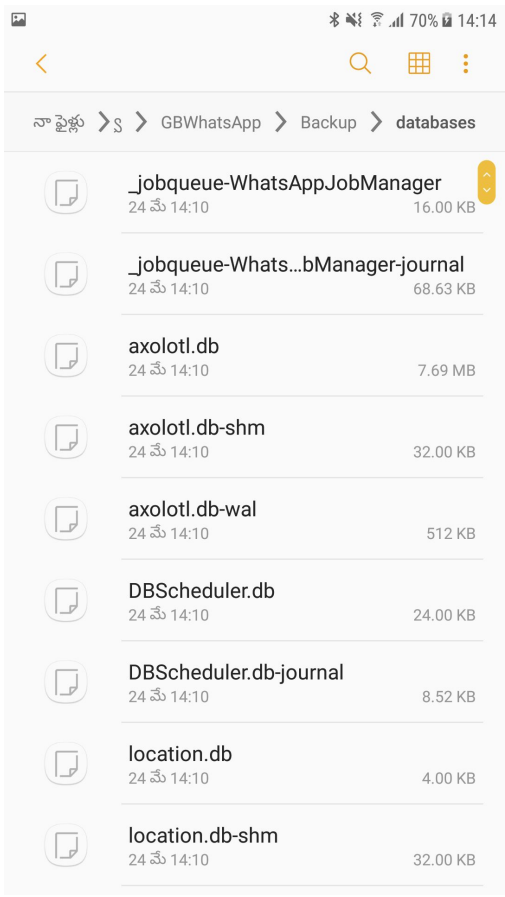


GB Settings -> Other MODS -> Backup Data


70% 14:14
నా ఫైళ్లు > అంతర్గత నిల్వ
18 మే 09:31 56 ఐటెమ్‌లు
Facebook Messenger
26 ఫిబ్ర 13:08 0 ఐటెమ్‌లు
FastScannerFree
19 ఏప్రి 19:19 5 ఐటెమ్‌లు
GBWhatsApp
24 మే 14:09 4 ఐటెమ్‌లు
gPodder
25 జూలై, 2017 16:44 1 ఐటెమ్
GROWTH_TEST
10 మే 09:22 1 ఐటెమ్
Kik
11 అక్టో, 2018 15:06 0 ఐటెమ్‌లు
motionstills
6 నవం, 2017 00:04 3 ఐటెమ్‌లు
MovieMakerLib
4 ఫిబ్ర, 2017 01:55 0 ఐటెమ్‌లు
Movies


86% 18:39
My Files > Internal storage > GBWhatsApp
Backup
24 May 14:10 5 items
Backups
12 Mar 02:00 2 items
Databases
27 May 02:08 9 items
Media
10 Mar 22:13 9 items


86% 18:39
My Files > nternal storage > GBWhatsApp >
cache
24 May 14:09 6 items
code_cache
24 May 14:09 0 items
databases
24 May 14:10 14 items
files
24 May 14:10 9 items
shared_prefs
24 May 14:09 15 items
Scroll capture
Draw
Crop
Share


70% 14:14
నా ఫైళ్లు > S > GBWhatsApp > Backup > databases
_jobqueue-WhatsAppJobManager
24 మే 14:10 16.00 KB
_jobqueue-Whats...bManager-journal
24 మే 14:10 68.63 KB
axolotl.db
24 మే 14:10 7.69 MB
axolotl.db-shm
24 మే 14:10 32.00 KB
axolotl.db-wal
24 మే 14:10 512 KB
DBScheduler.db
24 మే 14:10 24.00 KB
DBScheduler.db-journal
24 మే 14:10 8.52 KB
location.db
24 మే 14:10 4.00 KB
location.db-shm
24 మే 14:10 32.00 KB

What’s in this database?

Use e.g. DB browser for SQLite to view the database.

https://arxiv.org/pdf/1507.07739.pdf

Forensic Analysis of WhatsApp Messenger on Android Smartphones

Cosimo Anglano
DiSIT - Computer Science Institute,
Universitá del Piemonte Orientale, Alessandria (Italy)
email:cosimo.anglano@uniupo.it

This manuscript is the preprint of the paper *Forensic Analysis of WhatsApp Messenger on Android Smartphones*, that has been published on the *Digital Investigation* Journal, Vol. 11, No. 3, pp. 201–213, September 2014 doi:10.1016/j.diin.2014.04.003

(c) 2014. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/

Main Tables

chat_list

group_participants

messages

group_participants_history

Messages table

Field name	Meaning
_id	record sequence number
key_remote_jid	WhatsApp ID of the communication partner
key_id	unique message identifier
key_from_me	message direction: '0'=incoming, '1'=outgoing
status	message status: '0'=received, '4'=waiting on the central server, '5'=received by the destination, '6'=control message
timestamp	time of send if key_from_me='1', record insertion time otherwise (taken from the local device clock, and encoded as a 13-digits millisecond Unix epoch time)
received_timestamp	time of receipt (taken from the local device clock, and encoded as a 13-digits millisecond Unix epoch time) if key_from_me='0', '-1' otherwise
receipt_server_timestamp	time of receipt of the central server ack (taken from the local device clock, and encoded as a 13-digits millisecond Unix epoch time) if key_from_me='1', '-1' otherwise
receipt_device_timestamp	time of receipt of the recipient ack (taken from the local device clock, and encoded as a 13-digits millisecond Unix epoch time) if key_from_me='1', '-1' otherwise
send_timestamp	unused (always set to '-1')
needs_push	'2' if broadcast message, '0' otherwise
recipient_count	number of recipients (broadcast message)
remote_resource	ID of the sender (only for group chat messages)

Table 3: Structure of the *messages* table: fields storing message attributes

chat_list table

Contains the list of groups joined, and their names

group_participants table

Contains the list of participants in a group, along with their admin status.

group_participants_history table

- Contains information about events -- users joining, leaving, being kicked out of groups, users changing phone numbers, etc.
  - indicated by the ‘status’ field.

Few examples

Identifying users joining and leaving groups

Identifying images/videos

Identifying replies

Other interesting databases

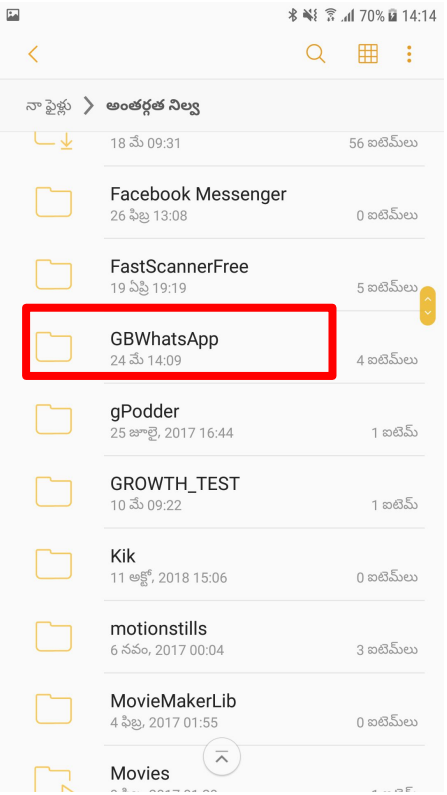
Contacts -- wa.db

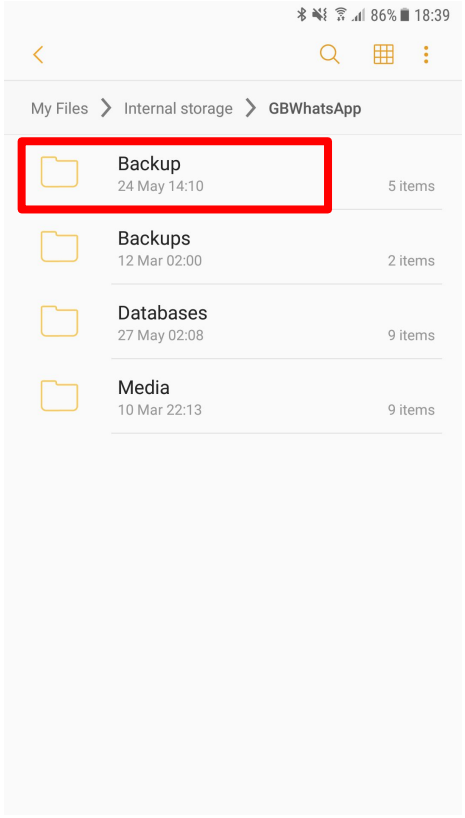
Keys -- axolotl.db

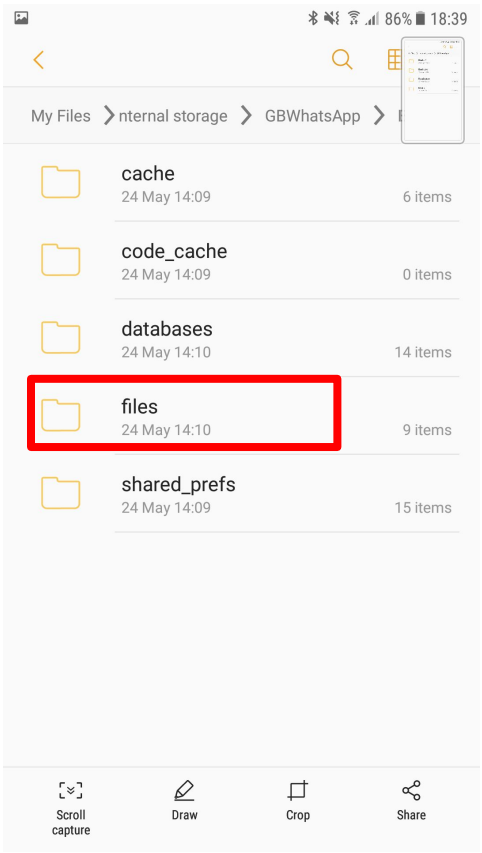
Also interesting: Logs

- Change of phone numbers
- Adding and deleting contacts
- Deleted messages

Logs location (jailbroken)







Chat reconstruction

- http://193.166.25.6/tmp/group_html_data/919481317437-1545561777@g.us.html

- You can click on the Toggle Sidebar button on the top to select from a random sample of 100 groups.
- Unfortunately it's hard to find a group completely in English. Some of the empty messages are videos which I did not including in the html.

- Code in the github repo task3/generateHTMLViewerFromDatabase.py

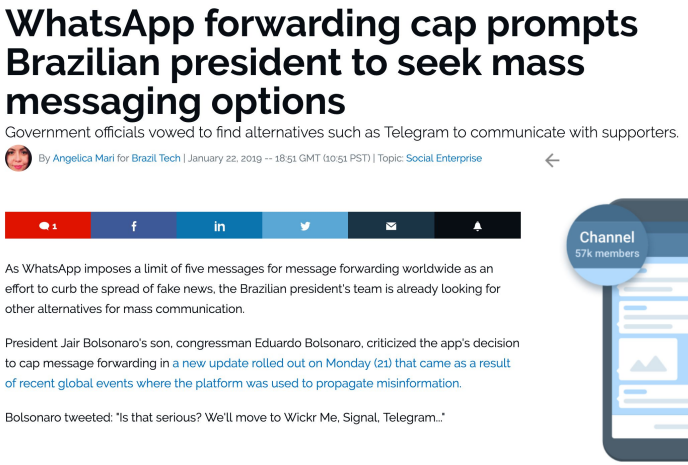
Other Messaging Platforms

- Other messaging platforms like Telegram, Kik, Discord, etc. have the same structure of storing data.
- Data stored as an sqlite3 database
- In /data/data

Quick look at Telegram

look for the unencrypted data in /data/data/com.telegram.com

closed but larger groups + channels



WhatsApp forwarding cap prompts Brazilian president to seek mass messaging options

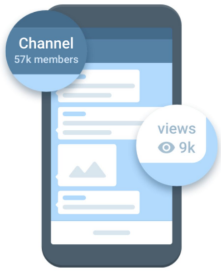
Government officials vowed to find alternatives such as Telegram to communicate with supporters.

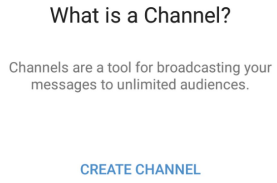
By Angelica Mari for Brazil Tech | January 22, 2019 -- 18:51 GMT (10:51 PST) | Topic: Social Enterprise

As WhatsApp imposes a limit of five messages for message forwarding worldwide as an effort to curb the spread of fake news, the Brazilian president's team is already looking for other alternatives for mass communication.

President Jair Bolsonaro's son, congressman Eduardo Bolsonaro, criticized the app's decision to cap message forwarding in a new update rolled out on Monday (21) that came as a result of recent global events where the platform was used to propagate misinformation.

Bolsonaro tweeted: "Is that serious? We'll move to Wickr Me, Signal, Telegram..."





Break

Let’s meet in 30 minutes.

Part 4 -- Case studies

What can be done with this data?

Agenda

- Potential problems

- Showcase (preliminary) work in this space
- Work from Brazil, India, Indonesia.
  - mostly politics
- Generalizable findings

Data collection 2

- Political groups

The New York Times

Opinion

Fake News Is Poisoning Brazilian Politics. WhatsApp Can Stop It.

By Cristina Tardáguila, Fabrício Benevenuto and Pablo Ortellado

Ms. Tardáguila, Mr. Benevenuto and Mr. Ortellado are the authors of a new report on misinformation in Brazil.

Oct. 17, 2018

Leer en español

FUTURE TENSE

Why Fake News on WhatsApp Is So Pernicious in Brazil

Brazilians use the app for *everything*. Bolsonaro and his supporters exploited that.

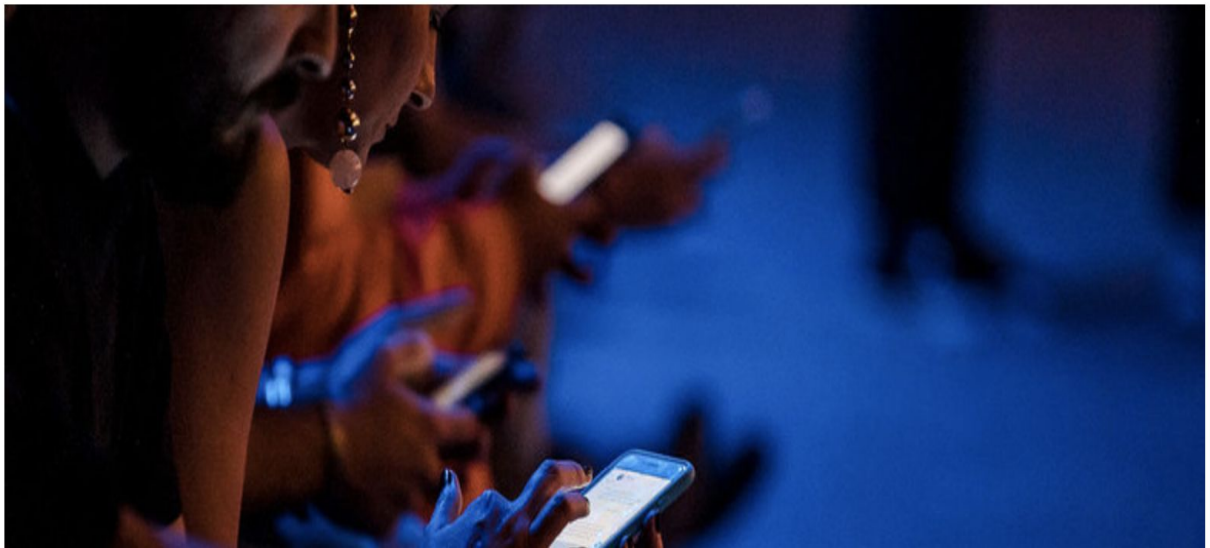
By TANIA MENAI

OCT 31, 2018 • 3:44 PM

TWEET

SHARE

COMMENT



Political groups

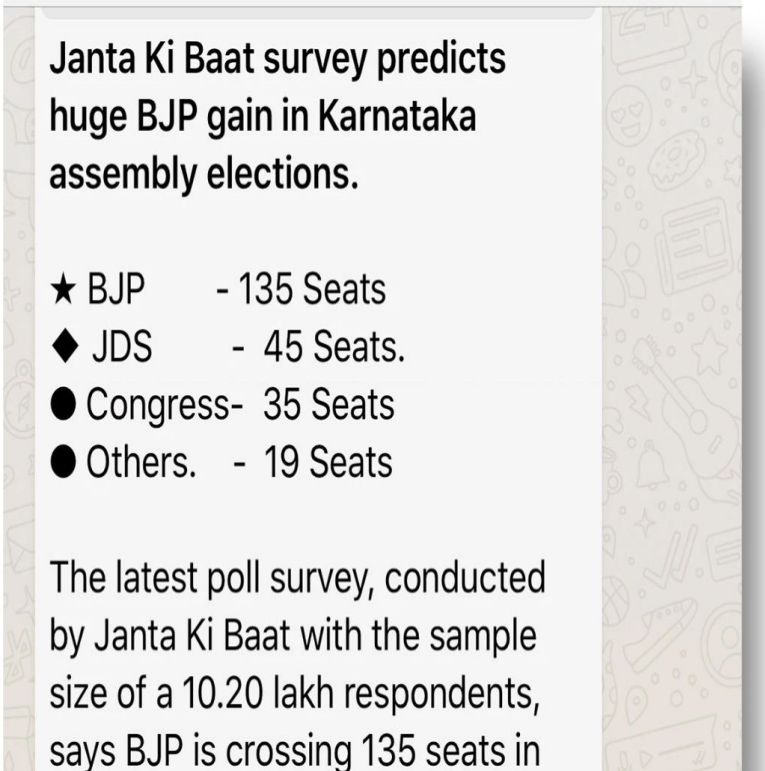
TECHNOLOGY

The New York Times

In India, Facebook's WhatsApp Plays Central Role in Elections

originated.

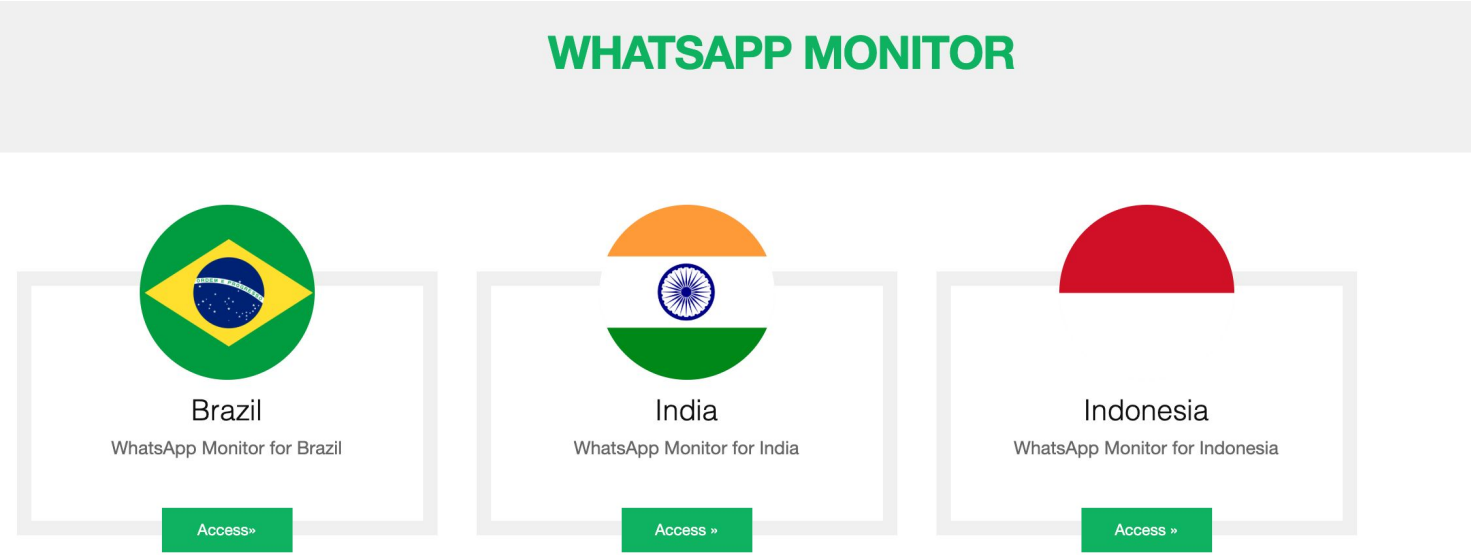
In the run-up to the May 12 vote in the state — the results of which are set to be announced on Tuesday — the B.J.P. and Congress parties claimed to have set up at least 50,000 WhatsApp groups between them to spread their messages. At the same time, many others — their identities are unknown — distributed videos, audio clips, posts and false articles designed specifically to rile up the area's Hindu-Muslim fissures.



They’re quite widely used!

- One in six WhatsApp users in India said they were members of a group started by a political leader or party

Political groups data



Political groups data

- We monitored a large number of political groups
- No idea about the representativeness of our data
- 35% of them images

- 10% of the messages sent by <1% of the users
- 10% of the groups created by <1% of the users

Potential directions

- Misinformation
- Political Coordination
- Effects on other platforms – YouTube, Twitter, FB, etc
- Education
- Other
  - Algorithmic vs. personal choices
  - Hate speech
  - External events
  - Cross cultural differences

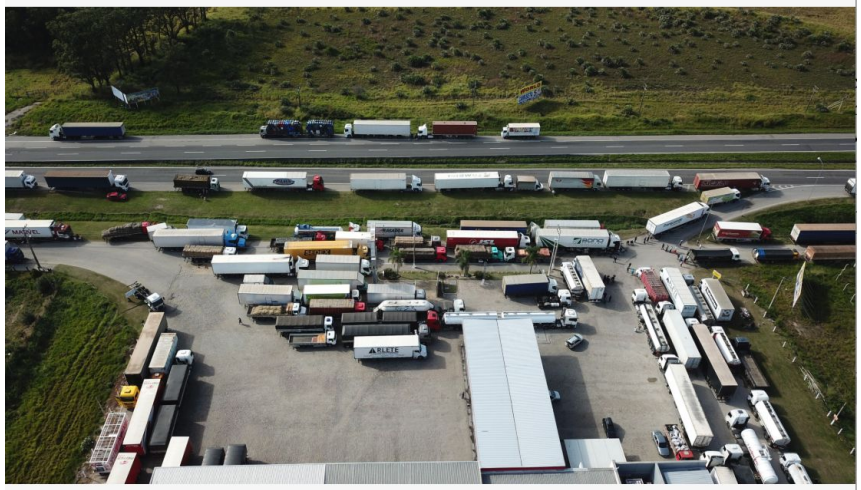
Misinformation

1. Examples from Brazil (WWW'19 paper, WebSci'19 paper)
2. Examples from India (much larger sample)

Events Captured in Brazil

ECONOMY & LABOR

The Brazilian Truckers' Strike: How WhatsApp Is Changing the Rules of the Game



Source:https://truthout.org/articles/the-brazilian-truckers-strike-how-whatsapp-is-changing-the-rules-of-the-game/

The New York Times

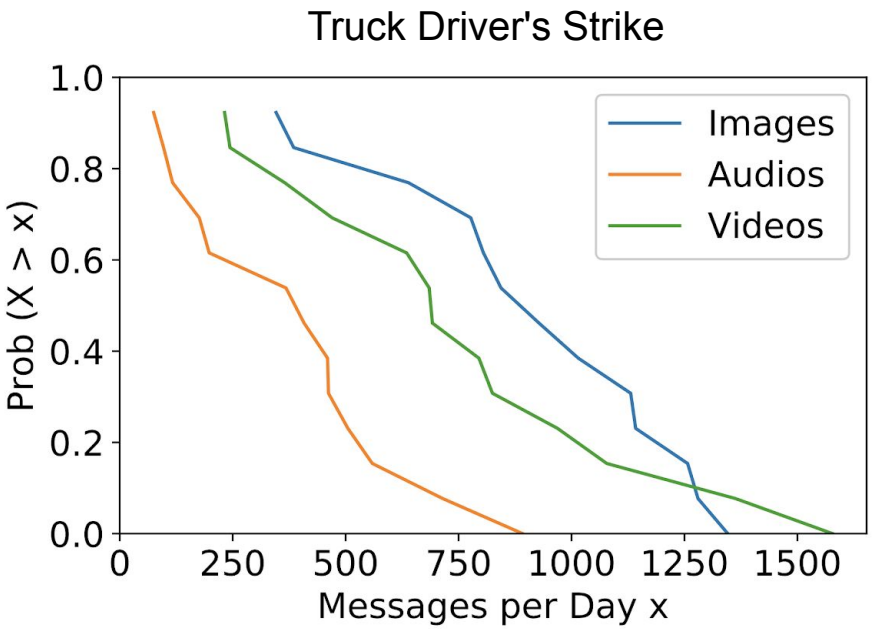
Disinformation Spreads on WhatsApp Ahead of Brazilian Election

Source:https://www.nytimes.com/2018/10/19/technology/whatsapp-brazil-presidential-election.html

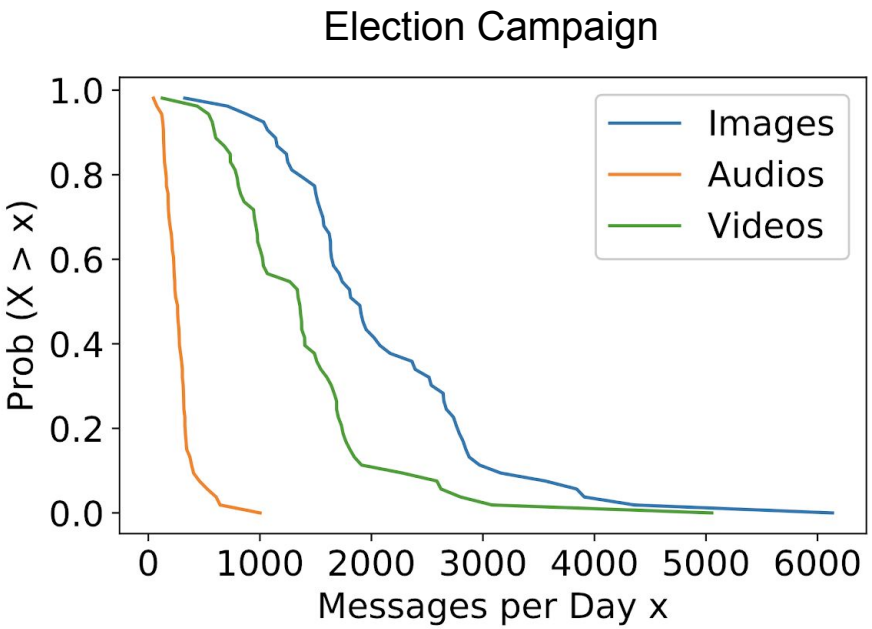
Table 1: Overview of our datasets.

	Truck Drivers' Strike	Election Campaign
#Groups	141	364
#Total Users	5,272	18,725
#Total Messages	121,781	789,914
#Text Messages	95,424	591,162
#Images	11,610	110,954
#Videos	9,752	73,310
#Audios	4,995	14,488
#URLs	11,728	92,654

Events Captured in Brazil



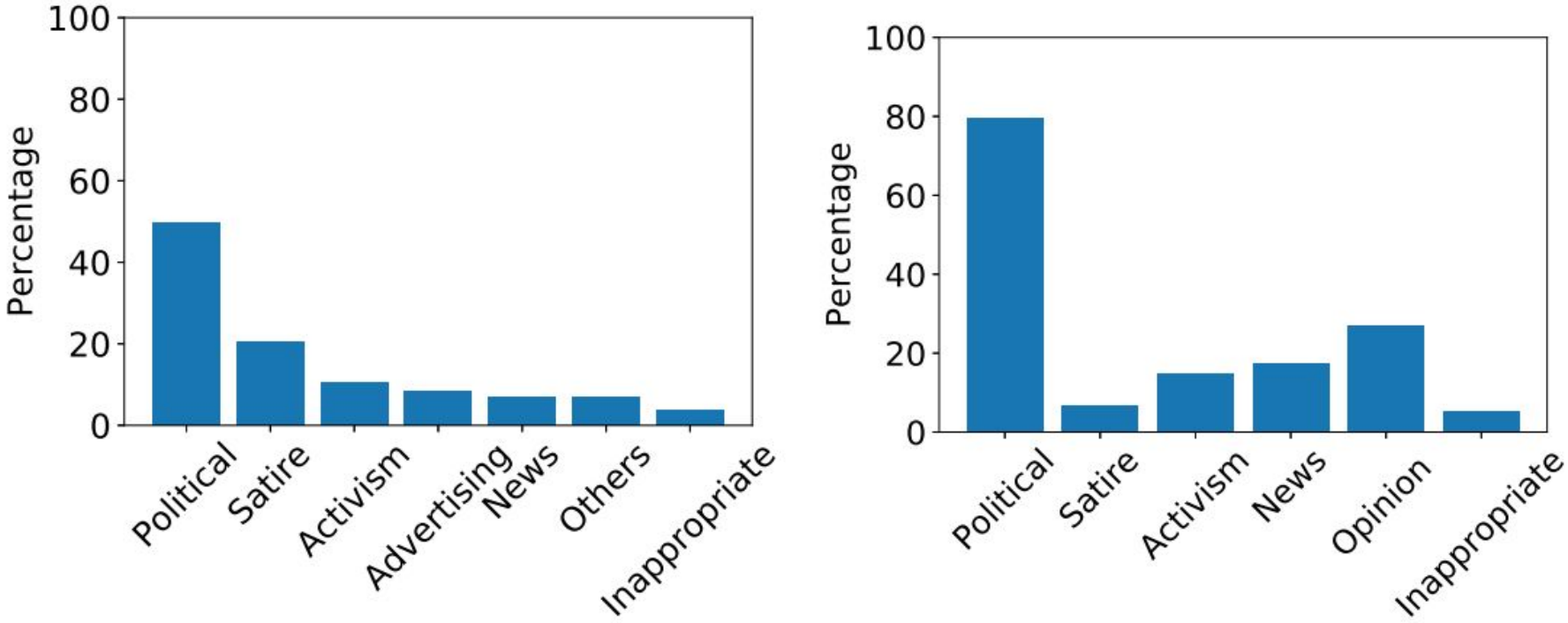
May 21st - Jun 2nd 2018



Aug 16st - Sep 7th 2018

Images: most popular media content shared on monitored groups

Manual annotation of images



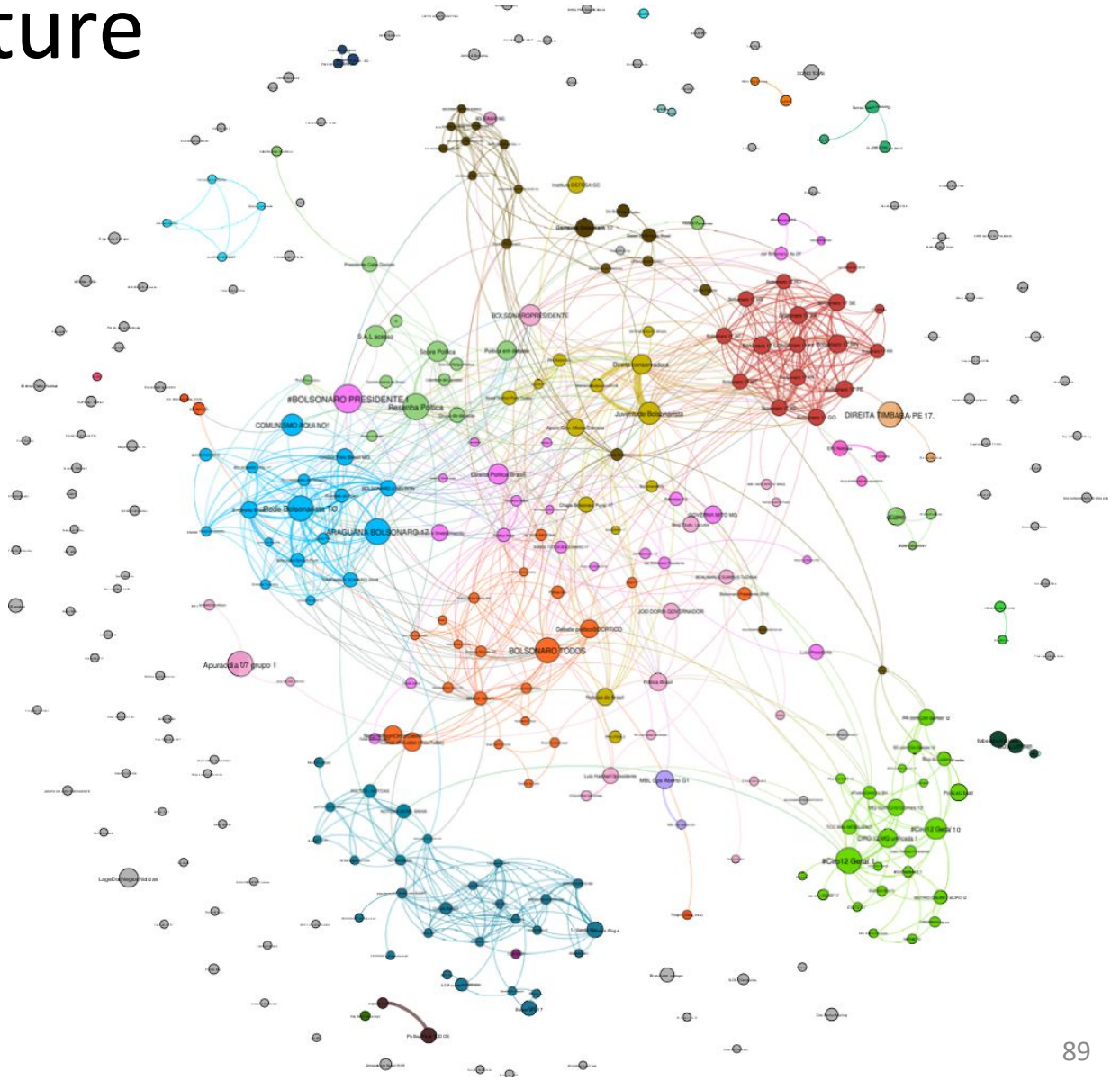
(a) Truck Drivers' Strike (b) Election Campaign

Figure 2: Distributions of image categories.

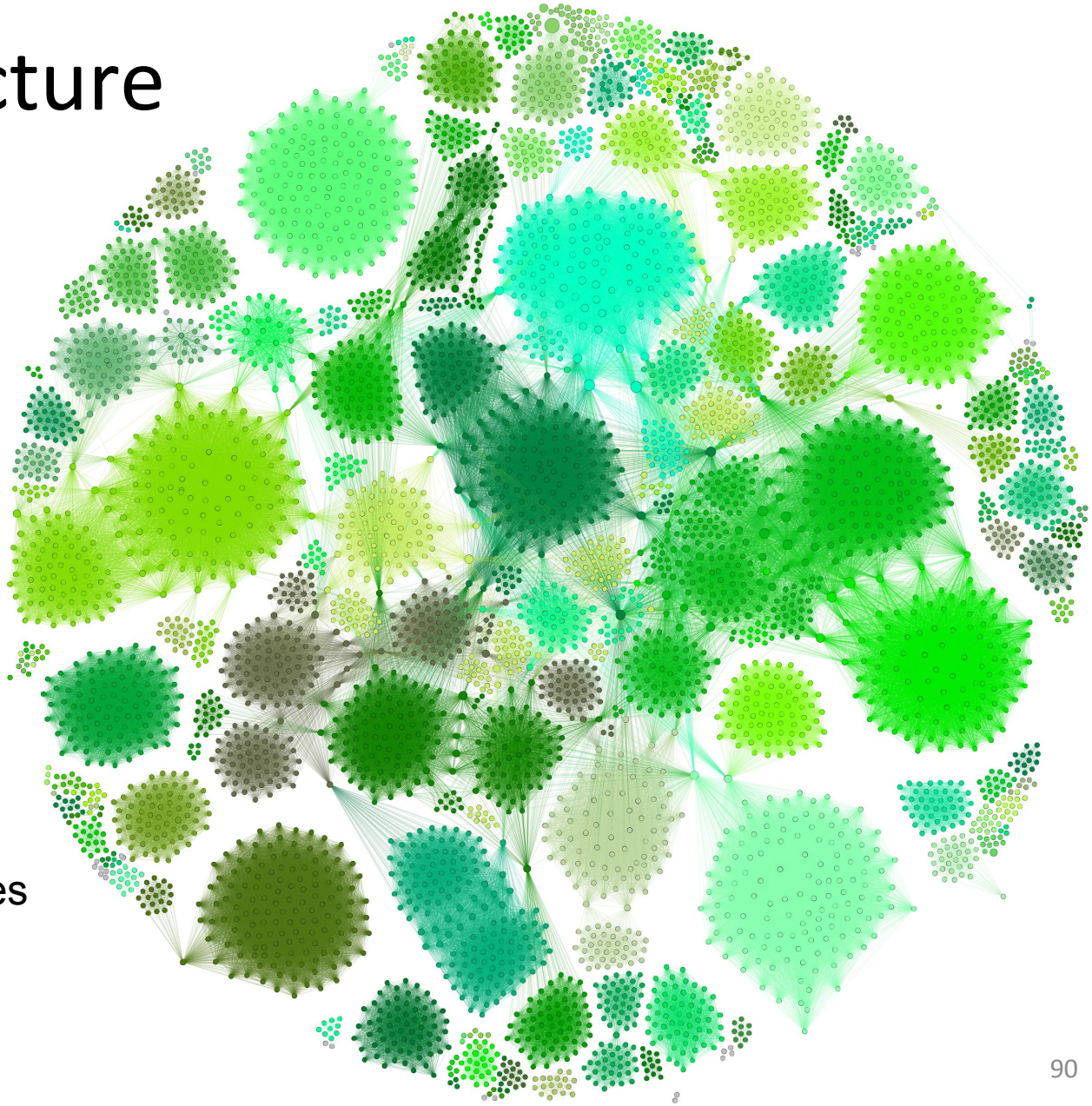
Network Structure

Group Network (election campaign)

node = group
edge = users in common
node size = # users



Network Structure



Users Network
(election campaign)

node = user
edge = groups common
node size = # groups user
shared images

Misinformation on WhatsApp

Labeling with a real Fact-Checking Agency

- Sample of 61 most shared images during Brazilian elections
- Checked by the Brazilian fact checking agency *Lupa*
- Images containing factual content
    - 36% contained misinformation
    - 53% include misleading and inconclusive content
    - 10% were verified as true

Misinformation on WhatsApp

An Automatic Methodology for Finding Misinformation

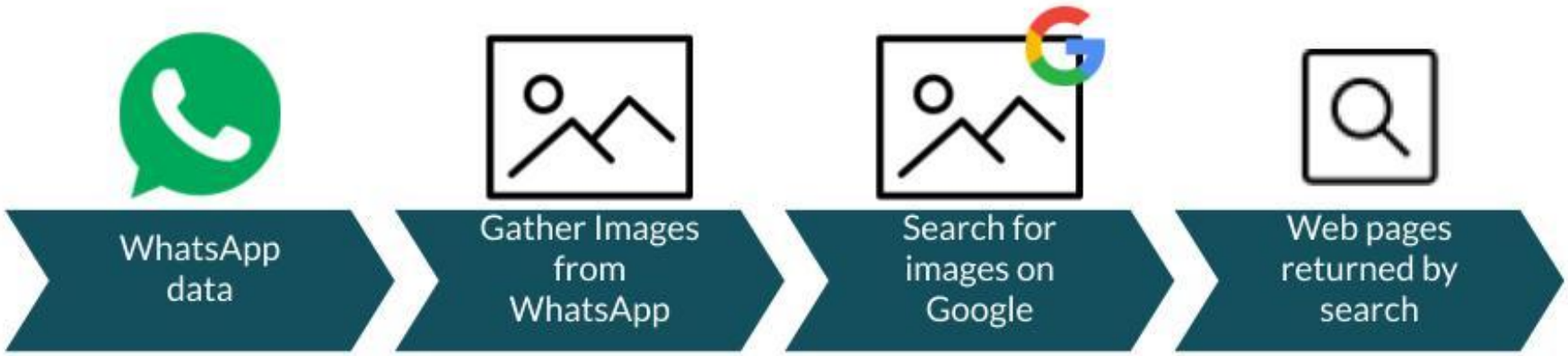
1. Generate list of main fact-checking agencies in the country
2. Automate search of each image from WhatsApp by using the Google Image search
3. Check whether any of the returned pages belong to one of the fact-checking domains
4. If so, parse fact-checking page and automatically label the image as fake or true

➢

Automatic identification of misinformation

- 70 images containing misinformation during election campaign
- Over 10k images identified for India

Collecting Sources of Images on the Web

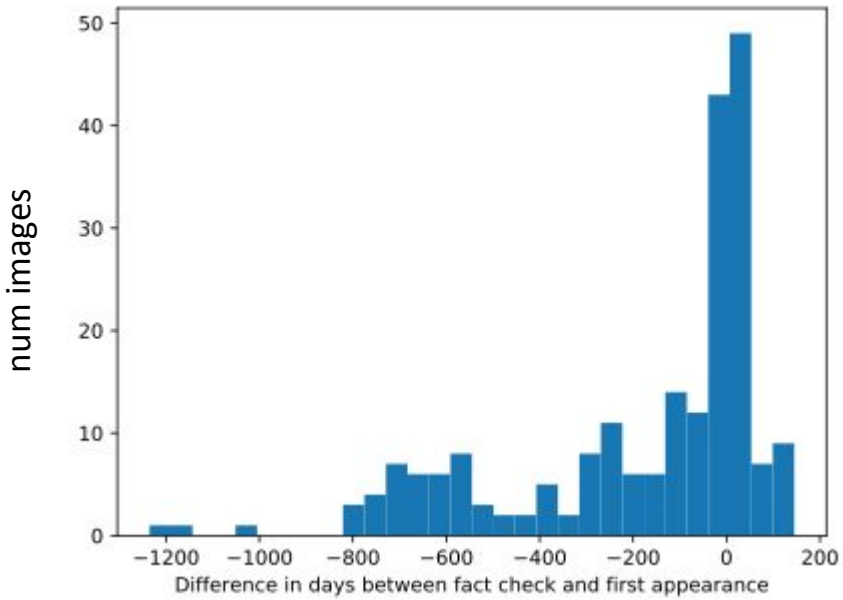


Search for collected images on Google Image search

Code included in the github repo.

Fact checked images

- 60% of the cases the images were fact checked before they appeared in our dataset

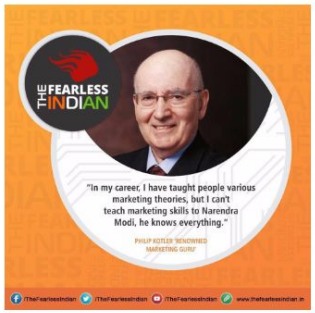


Examples

http://193.166.25.6/tmp/factcheck2.html (Caution: Graphic images!)

Same image shared again and again in different contexts.

Contains a sampling of the types of misinformation.



(a)

(b)

(c)

(d)

Figure 3: Examples of various types of misinformation images. (a) A false quote praising the Indian prime minister, (b) A false health scare, (c) A doctored image of a TV news program, and (d) An image of a plane crash taken out of context.

https://www.hindustantimes.com/india-news/decoding-fact-free-world-of-whatsapp/story-LQ79X96OOKrGo7MHuW3TMP.html

hindustantimes

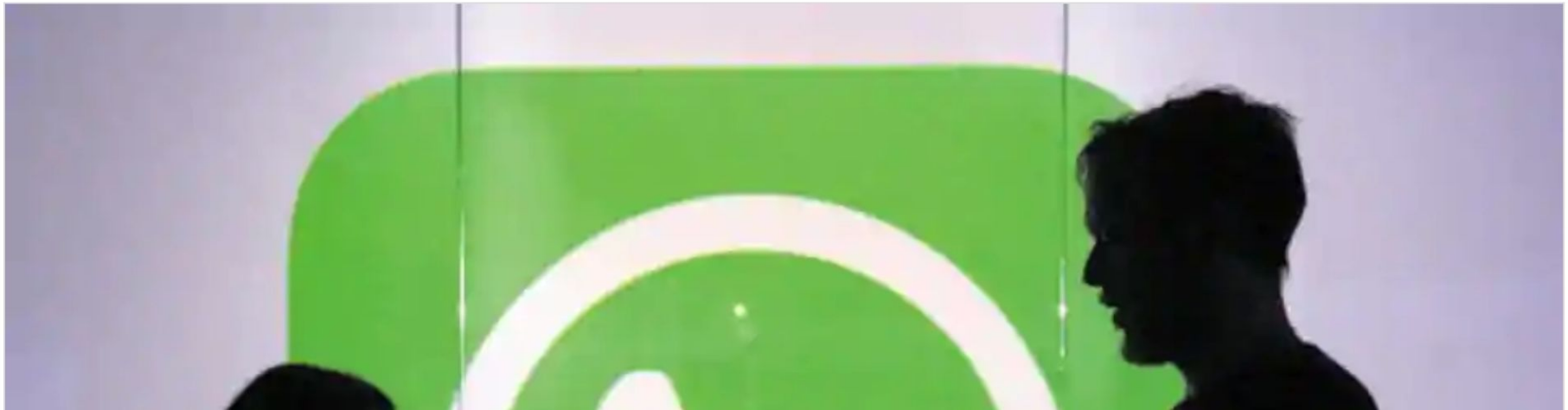
Tuesday, Mar 05, 2019

Fighting fake news: Decoding 'fact-free' world of WhatsApp

Misinformation Ecosystem: From outright falsehoods and misleading data to bigotry and hate, everything is circulated in the closed encrypted world of WhatsApp.

INDIA Updated: Mar 05, 2019 10:17 IST

Samarth Bansal and Kiran Garimella
Hindustan Times



The most shared fake images were doctored screenshots of TV news programs
Almost impossible to detect automatically

Clustering similar images

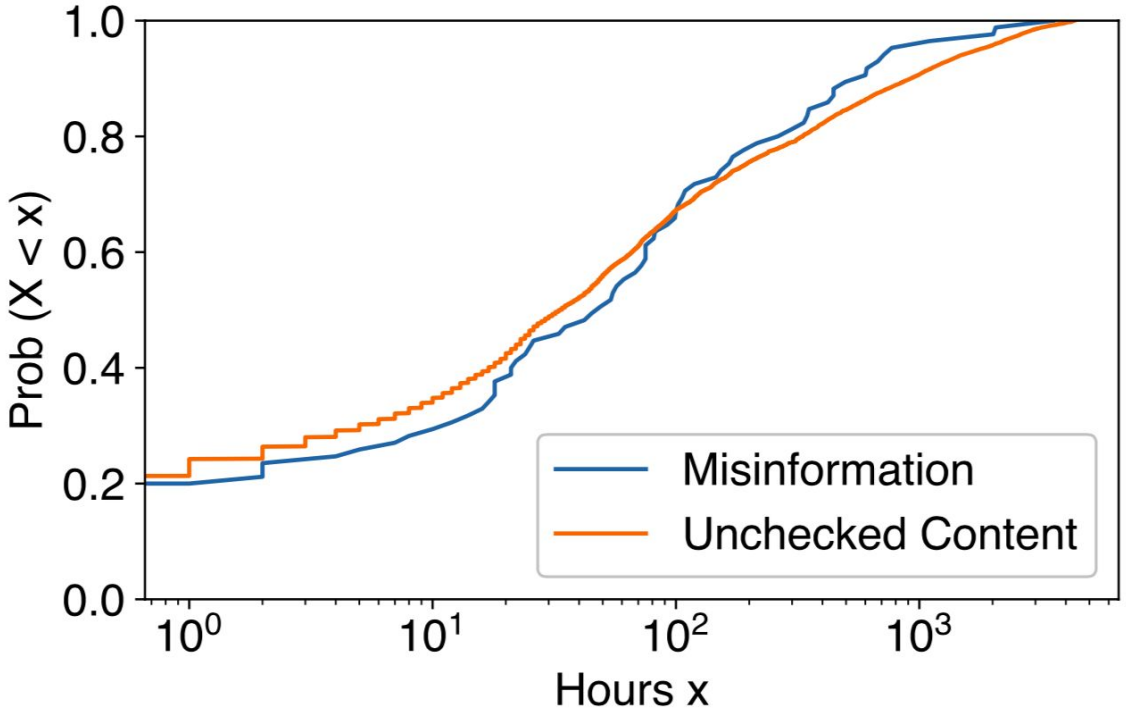
- Group similar images together using perceptual hashing tools, e.g.
  https://github.com/memespaper/memes_pipeline

- Misses a lot of stuff

Misinformation on WhatsApp

Propagation Dynamics:

LifeTimes = time difference between first and last sharing in any group



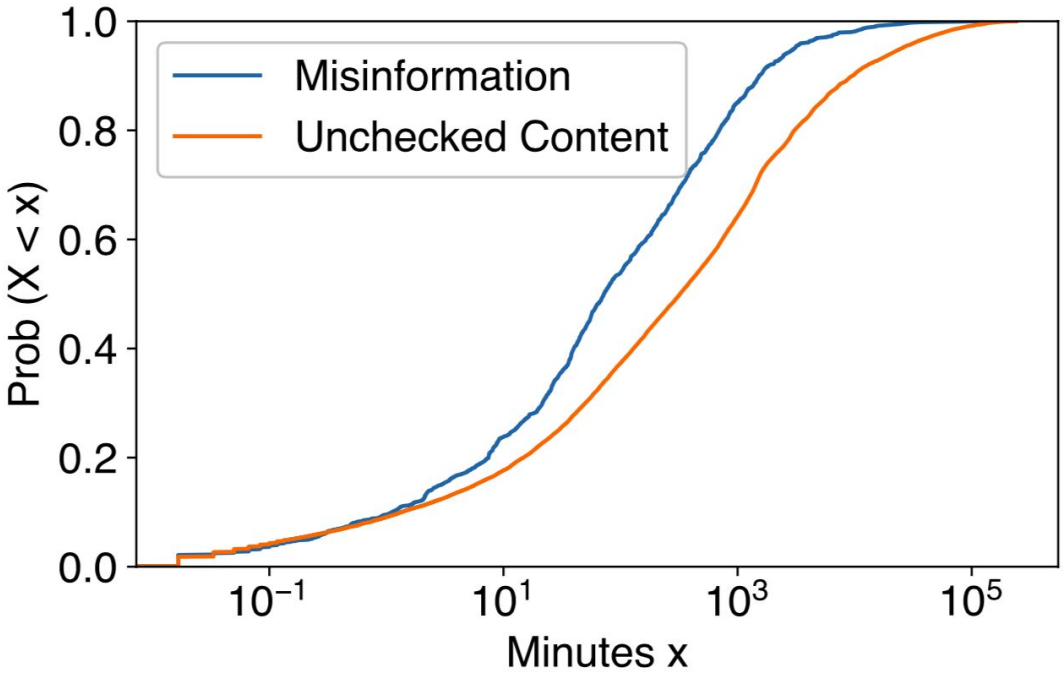
$$\Delta T = T_{last} - T_{first}$$

No significant difference, but some images remain for long time

Misinformation on WhatsApp

Propagation Dynamics:

Time difference between consecutive shares of same image in any group



$$\Delta T = T_{i+1} - T_i$$

Spread of images with misinformation is faster

Images with Misinformation and Unchecked Content

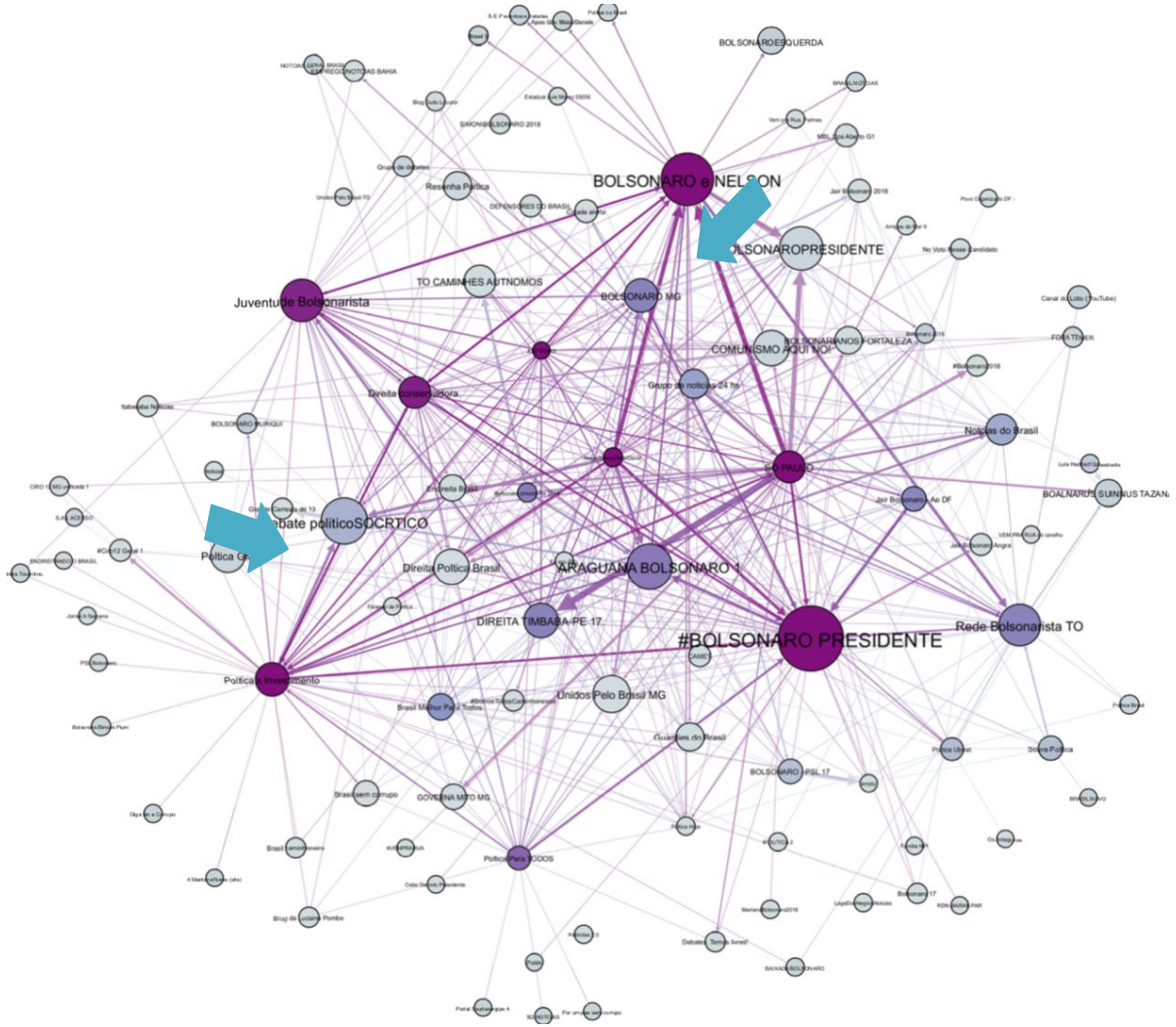
Election Campaign	Misinformation	Unchecked
#Groups in which images were shared	157	351
#Users who shared images	624	10,339
#Unique images	85	69,590
#Total shares of images	1,168	109,791

- Brazil: 85 images with misinformation; shared more than 1,000 times

Some peculiar aspects

- Mostly in-app
- No external links
- No bots (mostly)
- Malicious actors
- One broad solution not possible.

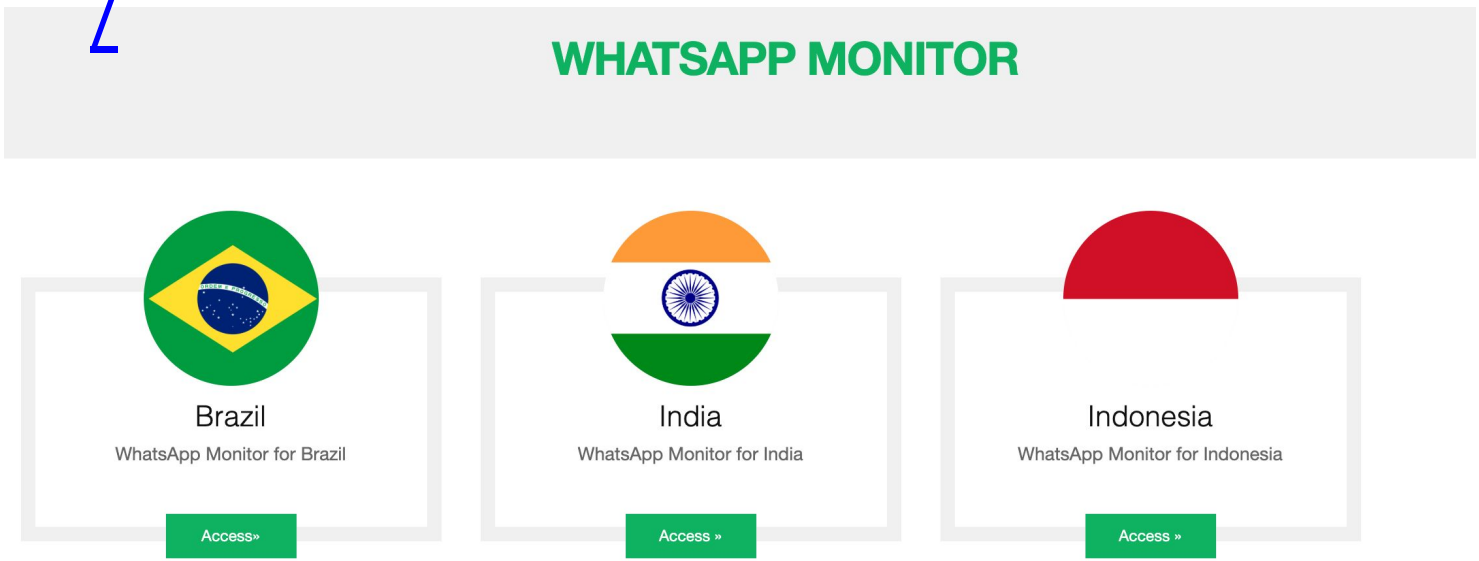
Misinformation Spread in WhatsApp



Large light nodes: mostly receptors, with many images shared

WhatsApp Monitor

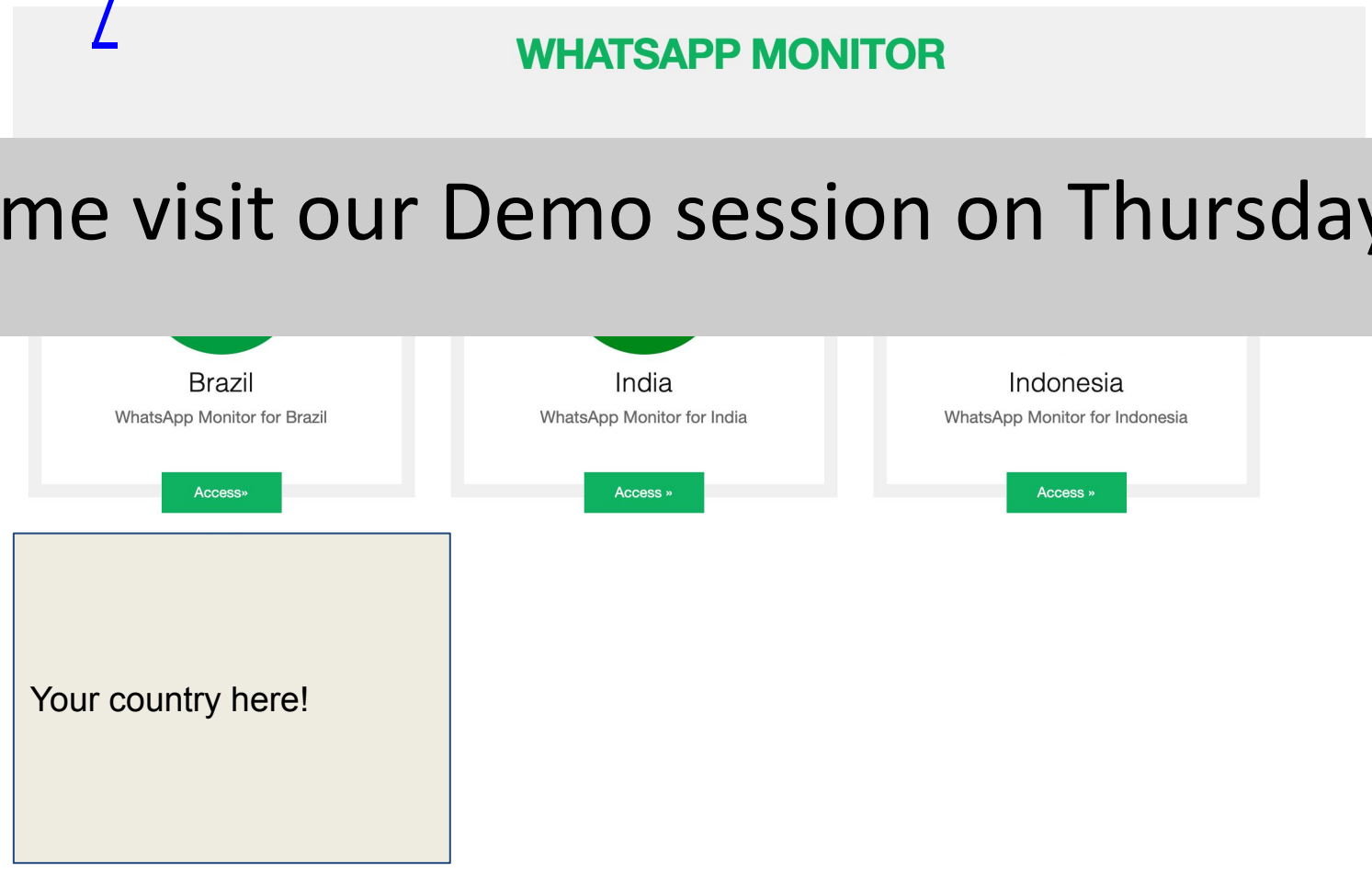
http://www.whatsapp-monitor.dcc.ufmg.br/



Your country here!

WhatsApp Monitor

http://www.whatsapp-monitor.dcc.ufmg.br/



Come visit our Demo session on Thursday!

WhatsApp Monitor

- Web based system
- Daily updated content
- Rank items by total shares on WhatsApp
- Content separated by type (images, messages, URLs, videos, audio)
- Restrict access through login and password

WhatsApp Monitor

Users can choose a date or interval of time

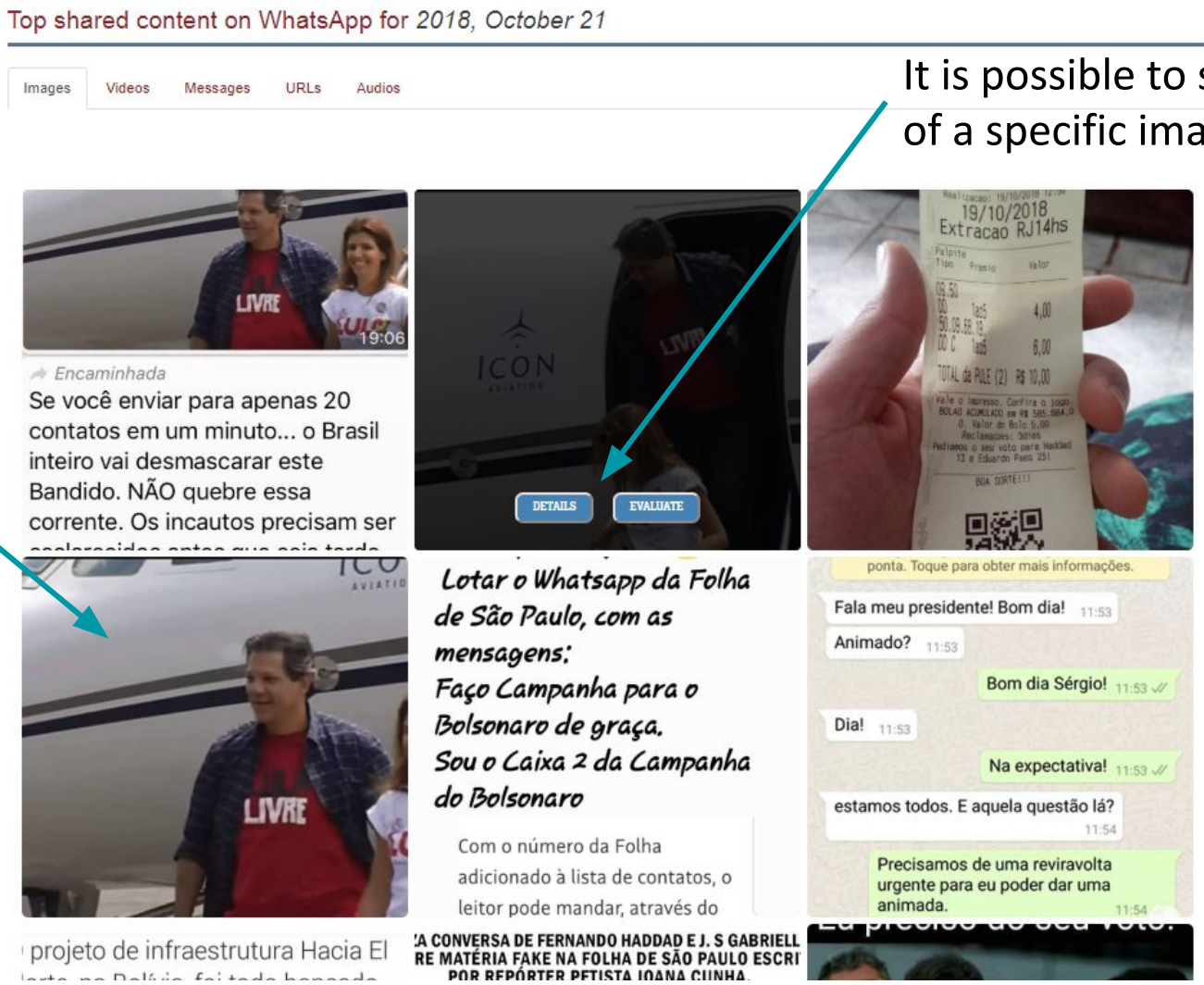
Navigate between days

Change type of content showed on system

WhatsApp Monitor

A list of images is showed for the date, ranked by total shares

It is possible to see details of a specific image



WhatsApp Monitor

For each content there is available details about amount of shares, number of different groups and different users that post them.

Image in bigger resolution

Propagation information about total of shares

List of groups where the image was shared

Search for this image on Web

Mark the image as false, NSFW, inappropriate content, etc

Messages before/after

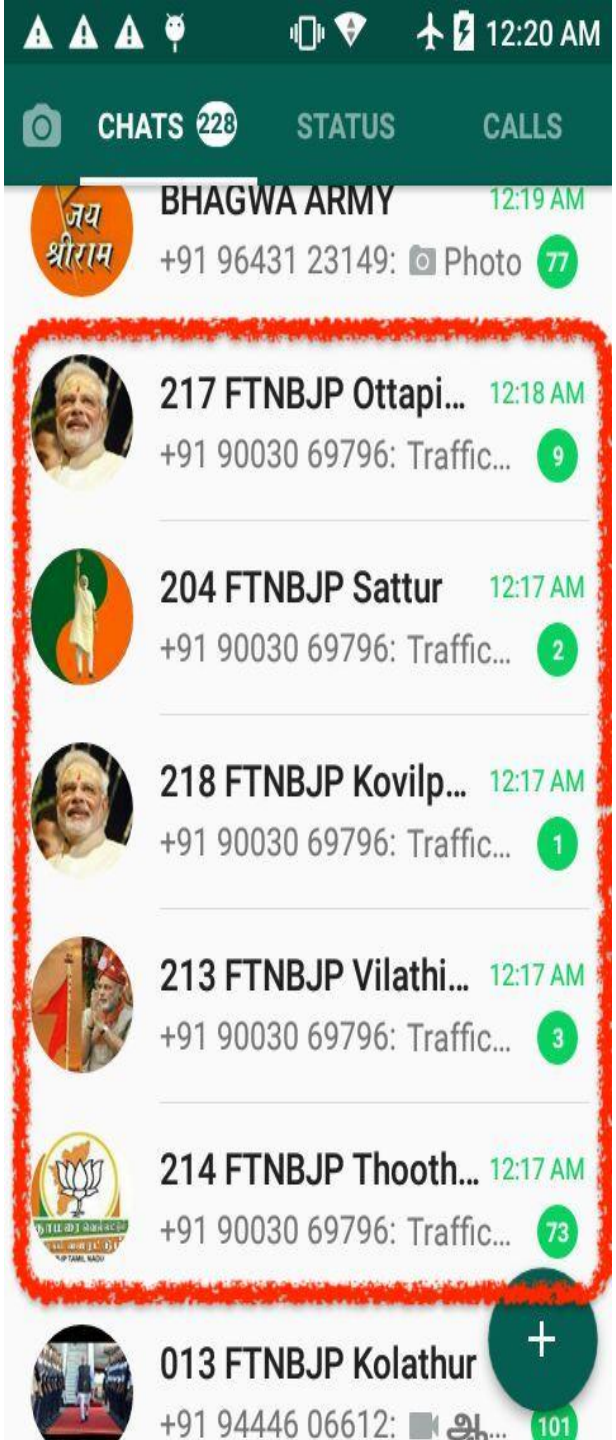
Central coordination

- How are political parties using WhatsApp to coordinate campaigns/messaging?

- Propaganda/Coordination is hard to track or be aware of the scale due to limited access.

Central coordination

- Over 10k cases of same message sent in >20 groups in a 10 second window.



Automation in the Spanish elections

News > World > Europe

Spain elections: WhatsApp suspends left-wing party Podemos's communication channel days before vote

The incident has sparked a debate about the role of tech giants in the political process

Jon Stone Europe Correspondent | @joncstone |
Thursday 25 April 2019 16:34 |

Like Click to follow The Independent

https://www.independent.co.uk/news/world/europe/spain-elections-whatsapp-podemos-channel-close-left-ing-de-olmo-a8886481.html

Case 2: Brazil university budget cuts

AP

Top Sto

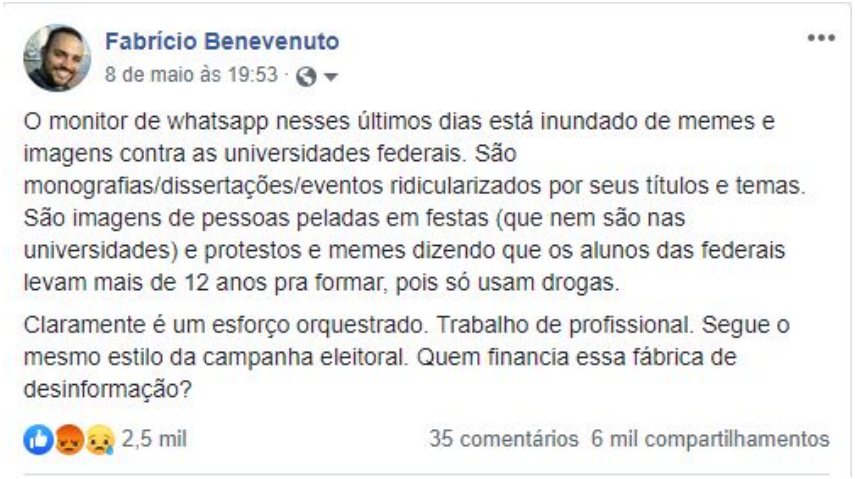
Brazil plans to slash funding of universities by 30 percent

By DIANE JEANTET May 1, 2019

Large scale micro targeting!

Images claiming to be from universities started spreading on WhatsApp groups.

Large scale micro targeting!



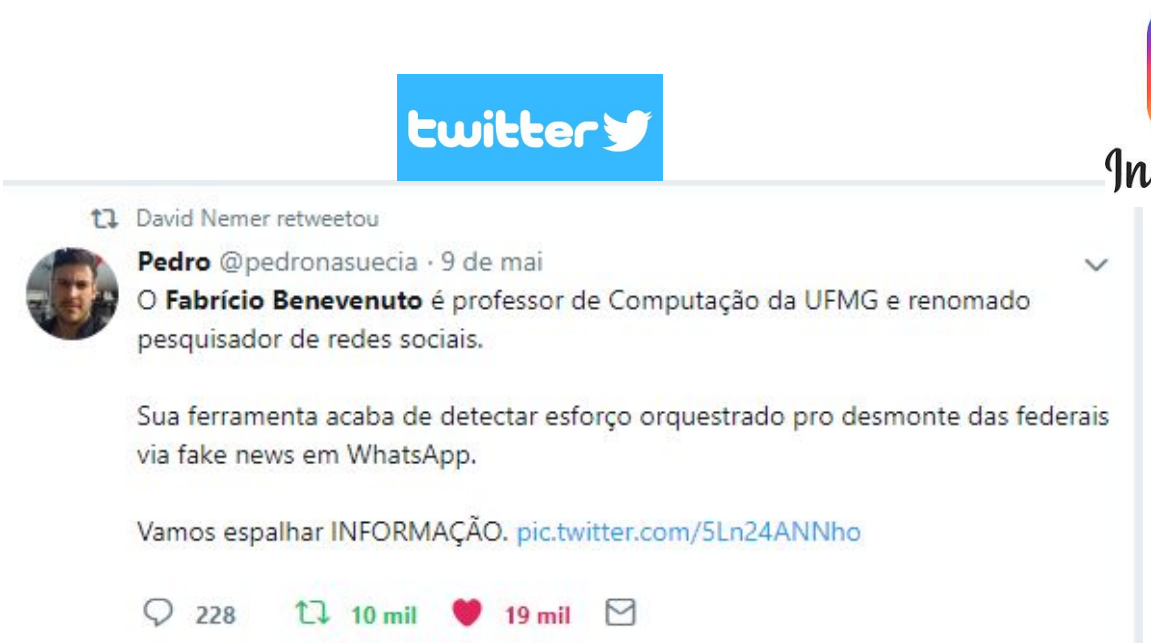
This happened just after a government cut of 30% in the universities funding

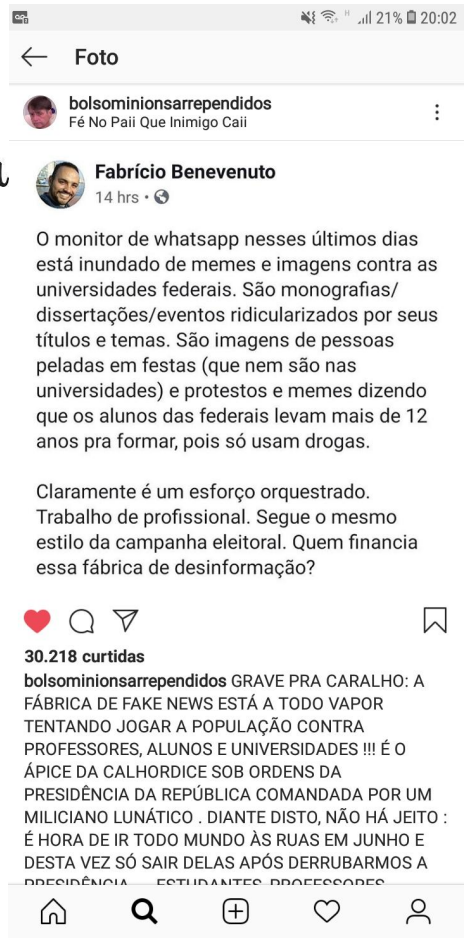
In these last days our WhatsApp monitor is flooded with memes and images against Universities. These images are about thesis mocked by their titles and themes and pictures of naked people inside universities and in parties. There are also memes saying students take 12 years to graduate because they only use drugs in universities.

This is clearly orchestrated and a professional work. Who is behind fabric of disinformation?

The Message Went Viral

Journalists with access to the system started to retweet and share.
Influencers and highly followed groups started to share it.

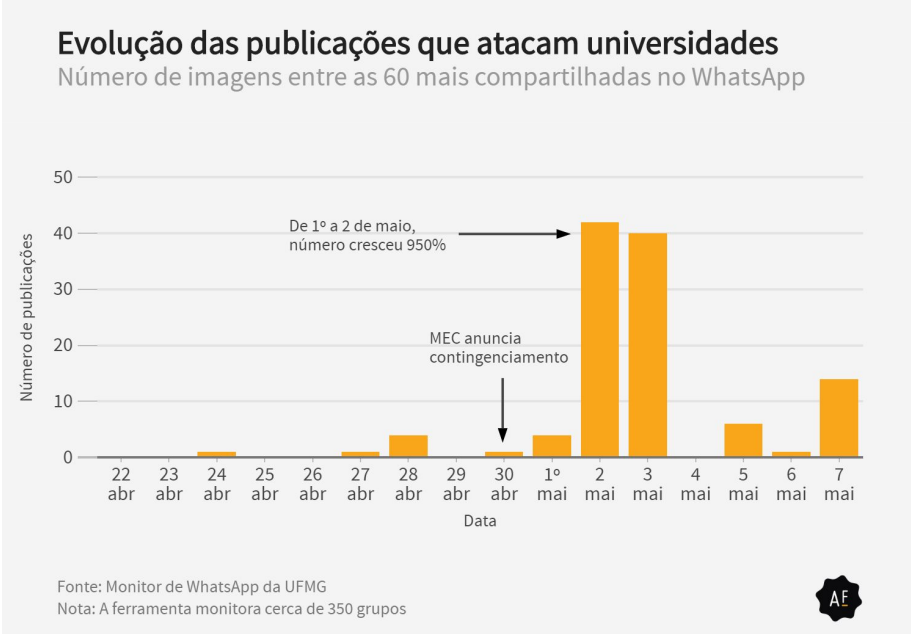




https://www.nationalgeographic.com/environment/2019/04/brazil-cuts-funding-scientists-grad-students-environment-suffers/

Quickly Fact checking agencies wrote about it

AOS FA-TOS



The increase perceived by Prof. Benevenuto was from 950% in comparison with the previous week

Lupa

A PRIMEIRA AGÊNCIA DE FACT-CHECKING DO BRASIL

Coordination on Twitter

- Google docs containing list of tweets
- Hashtag to get trending.
- Over 80% trending

#RaGaFailOnRafale
09.30 AM
15.11.2018

"We have lots of advantages in the Rafale deal...this and the S400 will enhance capabilities to counter our shortfalls." - Chief of the Air Staff, Air Chief Marshal BS Dhanoa //economictimes.indiatimes.com/articleshow/66051349.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst #RaGaFailOnRafale

"It is for the government to choose. It was decided to buy two squadrons through a Government to Government deal, to meet up emergency requirements." Chief of the Air Staff, Air Chief Marshal BS Dhanoa exposes Rahul Gandhi's lie. //economictimes.indiatimes.com/articleshow/66051349.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst #RaGaFailOnRafale

Rahul Gandhi must be really stupid to call the CEO of an internationally reputed company a co-accused in the fake Rafale Scam. #RaGaFailOnRafale

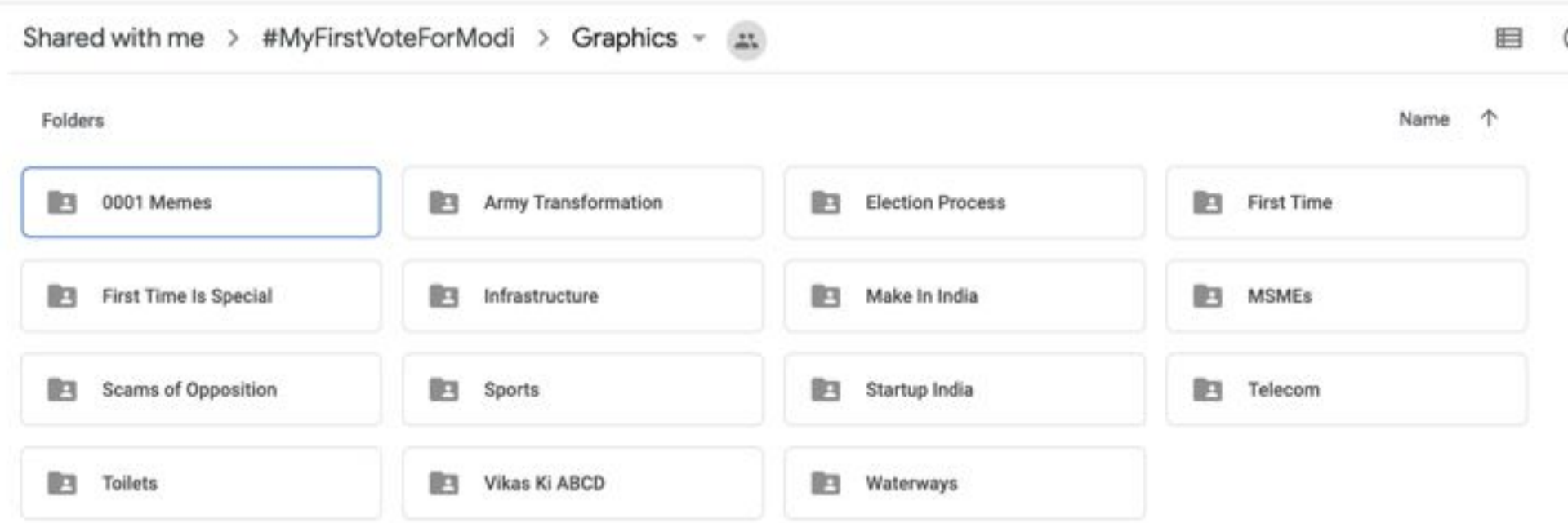
"It's a very good aircraft. When it comes to the subcontinent, it will be a game changer, and we have lots of advantages in the Rafale deal, we have got a very good package. It has got the best sensors, advanced weaponry." -Chief of the Air Staff. Still Rahul Gandhi wants to sabotage the Rafale Deal. #RaGaFailOnRafale //economictimes.indiatimes.com/articleshow/66051349.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst

From Dassault's CEO to experts in France, everybody has exposed the lies of Rahul Gandhi and Congress people, still they won't quit lying and hurting our national security. #RaGaFailOnRafale

It is so shameful that Rahul Gandhi is dragging the name of a reputed company into the petty politics of his party. At least don't hurt business interests of an international0 company for your political interests, Mr. Gandhi? #RaGaFailOnRafale

Modi got Yoga recognised at international level with International Yoga Day. Now countries in every corner of the world are organising mass yoga events. Finally due to efforts of BJP government, Hindu traditions are being recognised world over. #TakingForwardHinduCause

It’s well planned

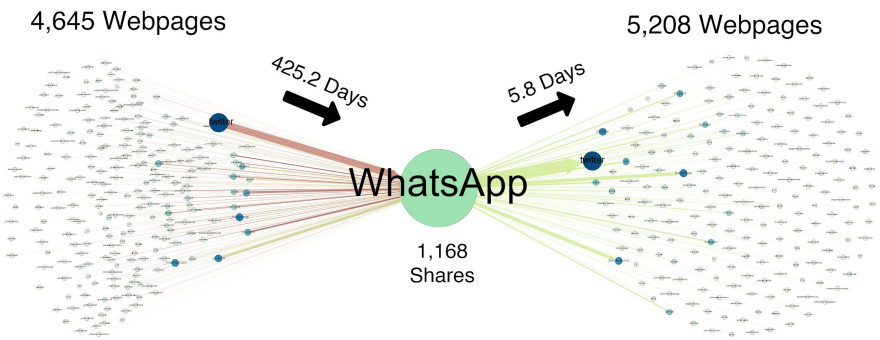


Effects on other platforms

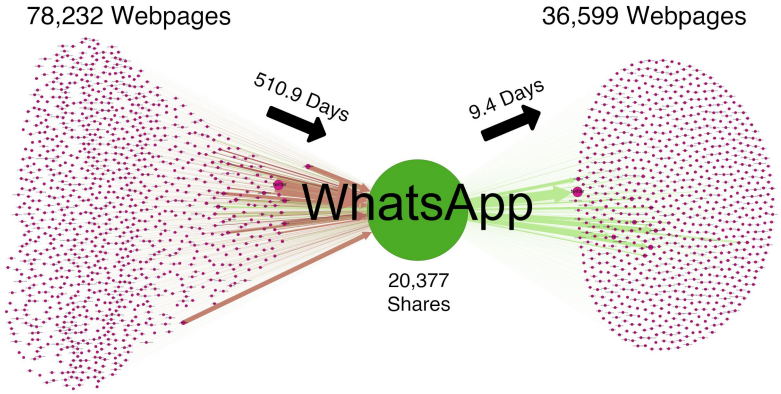
- Where does misinformation come from?
- What role does WhatsApp play in this?

Temporal Ordering of First Appearance: Web vs. WhatsApp Groups

Images with misinformation



Images with unchecked content



WhatsApp was the primary source of 30% of the identified images containing misinformation

Images with misinformation much more often appear first on WhatsApp and then on the Web

Effects on other platforms

Namo app

Sharechat

Tiktok

WhatsApp

Ads

Twitter

Youtube

Facebook

Mainstream media

Spread of information across platforms

whatsapp 2019-04-16 20:41:37

The Boston

Sunday, September 30, 2001 page13

Indian Politician Arrested

An Indian politician was detained at Boston airport when the airport security found him in possession of banned drugs and unaccounted cash. As per our sources, he is the son of a former Indian Prime Minister. He was later released after the intervention of the Indian ambassador to the US, reports AFP.

facebook 2019-04-21 13:26:04

The Boston

Sunday, September 30, 2001 page13

Indian Politician Arrested

An Indian politician was detained at Boston airport when the airport security found him in possession of banned drugs and unaccounted cash. As per our sources, he is the son of a former Indian Prime Minister. He was later released after the intervention of the Indian ambassador to the US, reports AFP.

sharechat 2019-04-19 09:26:10

The Boston

Sunday, September 30, 2001 page13

Indian Politician Arrested

An Indian politician was detained at Boston airport when the airport security found him in possession of banned drugs and unaccounted cash. As per our sources, he is the son of a former Indian Prime Minister. He was later released after the intervention of the Indian ambassador to the US, reports AFP.

Comparison with other platforms

- Flow of information
  - what type of memes originate where
  - closed vs. open platforms

- Quantifying algorithmic vs. personal choices

Not just large scale studies

- WhatsApp as a survey and data collection tool for field research

WhatsApp for public Education

- In South Africa, USAID used WhatsApp to **codesign educational campaigns** for the dapivirine ring to reduce HIV infections.
- In Zambia, **IntraHealth International** used WhatsApp to help rural nurses and midwives **connect with each other** for support and mentoring day and night.
- In Brazil, doctors used WhatsApp to **track Zika symptoms** and discover its link to microcephaly birth defects.

https://www.intrahealth.org/vital/ready-whatsapp-revolution-health-care-communications

Challenges

- Multimedia data and misinformation
  - video/audio are out of scope, without humans
  - tools for images are not mature enough
- Labelling issues
  - Context
  - Ethics
  - Malicious intent
- Text problems
  - Context
  - Local Languages

Challenges

- Mitigating misinformation
  - closed platform

Misinformation in multiple modalities

This image in itself does not mean anything.

Potential directions

- Misinformation
- Central Coordination
- Effects on other platforms – YouTube, Twitter, FB, etc
- Effects of forwarding restrictions
- Quantitative analysis
  - Hate speech
  - External events
  - Algorithmic vs. personal choices
  - Cross cultural differences

Can be generalized

- Ongoing data collection:
  - Uber-like bike riders in Indonesia
  - Syrian refugees in Turkey
  - Alumni groups from Indonesia
- Same for Telegram
  - Pro and anti government groups in Iran

Limitations

- Only ~10% of the conversations on WhatsApp are through groups
- Limited context – what happens outside the groups is not visible
- Sampling biases
- Hard to scale

Other issues

https://techcrunch.com/2018/12/20/whatsapp-pornography/

illegal refugee smuggling

local businesses

immigrant communities

Illegal refugee smuggling

- “Smuggling to Europe”
- "Migration to Europe"
- "Migration from Turkey Greece "
- "Crossing into Turkey and Europe”
- "Reversible Yunnan Turkey"

Translated from Arabic. Groups obtained by searching for ‘smuggle’ in Arabic.

Smuggling to Europe,❤️❤️

[redacted]

The way of reclining

4:58 AM

[redacted]

~لو كان صديقا

Ok I am married and I have 4 children about؟*

4:59 AM

I mean 6 people, my wife and 4 children

4:59 AM

How much is the cost per one

5:00 AM

[redacted]

?What brother

5:00 AM

Each win is 1400 euros

5:00 AM

Continue to Thessaloniki

5:00 AM

[redacted]

~لو كان صديقا

I mean how much is in dollars

5:01 AM

[redacted]

1600 $

5:01 AM

[redacted]

~

I mean 16 sheets

5:02 AM

Right

5:02 AM

[redacted]

Yes, my brother

Discussion

- Ethics -- public groups, but contain private phone numbers. What do we do?
- Contains illegal activity -- what do we do?
- Releasing the data for research?
- Can be easily misused
  - personal examples of malicious actors

Discussion

- Tools for processing whatsapp data
  - images, local language text
- Using whatsapp as a survey tool
- Mitigating misinformation -- what do we do?
- Any other use cases come to mind?

Acknowledgements

- Dean Eckles, MIT (Research support)
- Tarun Chitta, IIIT NR (Data collection)
- Fadlan Anam (Indonesia data)
- Numerous other forums and blogs

Bibliography

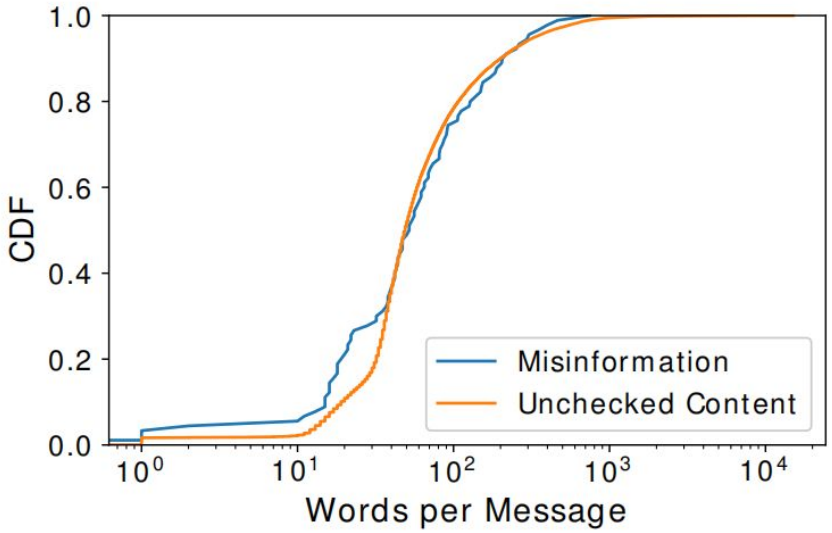
1. WhatsApp Doc?: A First Look at WhatsApp Public Group Data (ICWSM 2018)
2. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C22&q=wa.db+whatsapp&btnG= (list of related work analyzing whatsapp database)
3. Forensic Analysis of WhatsApp Messenger on Android Smartphones https://arxiv.org/pdf/1507.07739.pdf
4. Characterizing Attention Cascades in WhatsApp Groups https://arxiv.org/abs/1905.00825
5. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures (WWW 2019)
6. Analyzing Textual (Mis)Information Shared in WhatsApp Groups (WebSci 2019)
7. Forensic analysis of encrypted instant messaging applications on Android

Thank you!

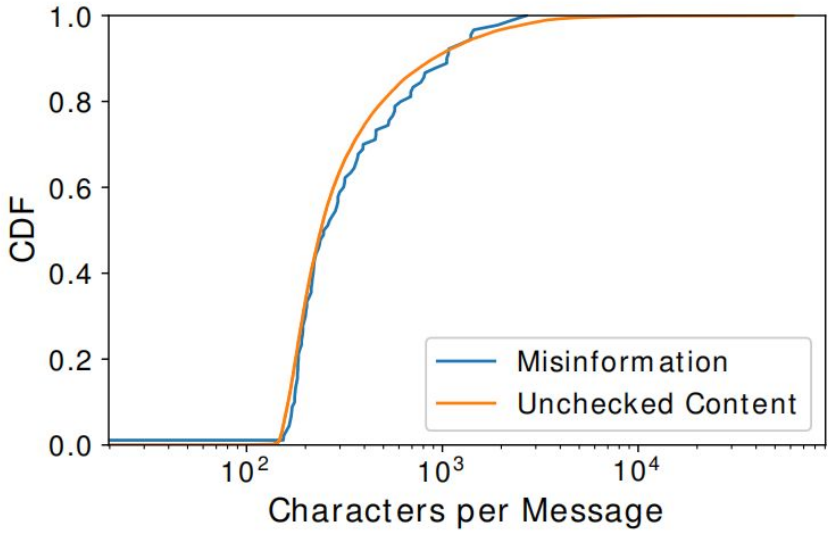
Additional slides

WHATSAPP WEBSCI

Distributions of message sizes:



(a) Number of words



(b) Number of characters

WHATSAPP WEBSCI

Word Cloud of the top 500 words

(a) Messages with misinformation

WHATSAPP WEBSCI

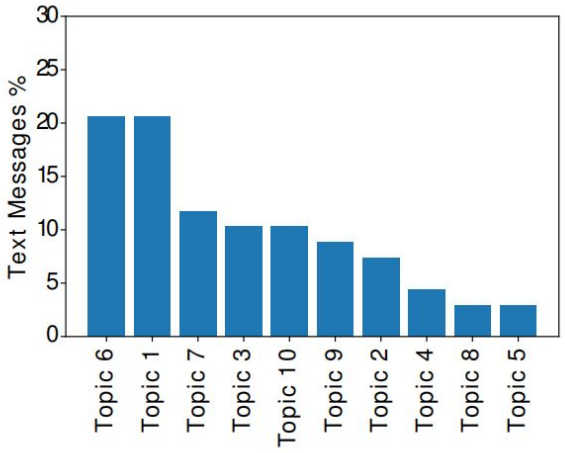
Word Cloud of the top 500 words

(b) Messages with unchecked content

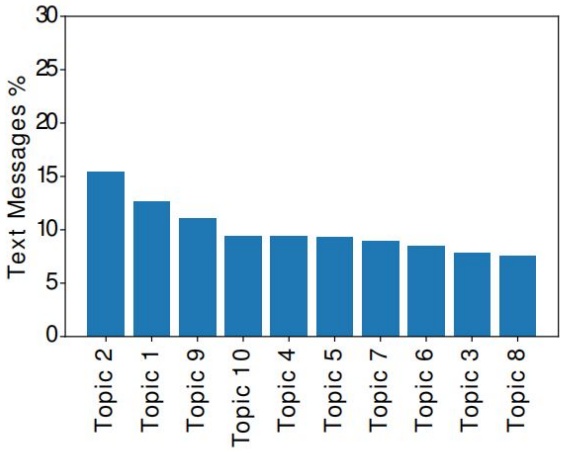
WHATSAPP WEBSCI

Topics inferred by LDA algorithm and distributions of topics inferred by LDA

Topic	Most representative words (translated to English)
1	vote, president, Haddad, Lula, Ciro, apply, research, PT, elections, voter
2	no, ant, know, do, person, speak, find, thing, expensive, people
3	say, life, God, do, Lord, day, man, no, good, be
4	country, nation, Brazilian, Brazil, left-wing, political, power, party, govern, right-wing
5	be, laugh, city, governor, senator, yes, state federal, new, big
6	govern, money, do, work, company, millions, year, Brazilian real, pay, receive
7	Bolsonaro, Brazil, say, woman, support, Jair, defend, apply, see, favor
8	be, law, publish, form, education, leave, be, use, project, right, project
9	day, group, Facebook, video, today, folks, chat_whatsapp, friend, share, hour
10	year, cop, after, weapon, news, city, arrested, find, crime, where



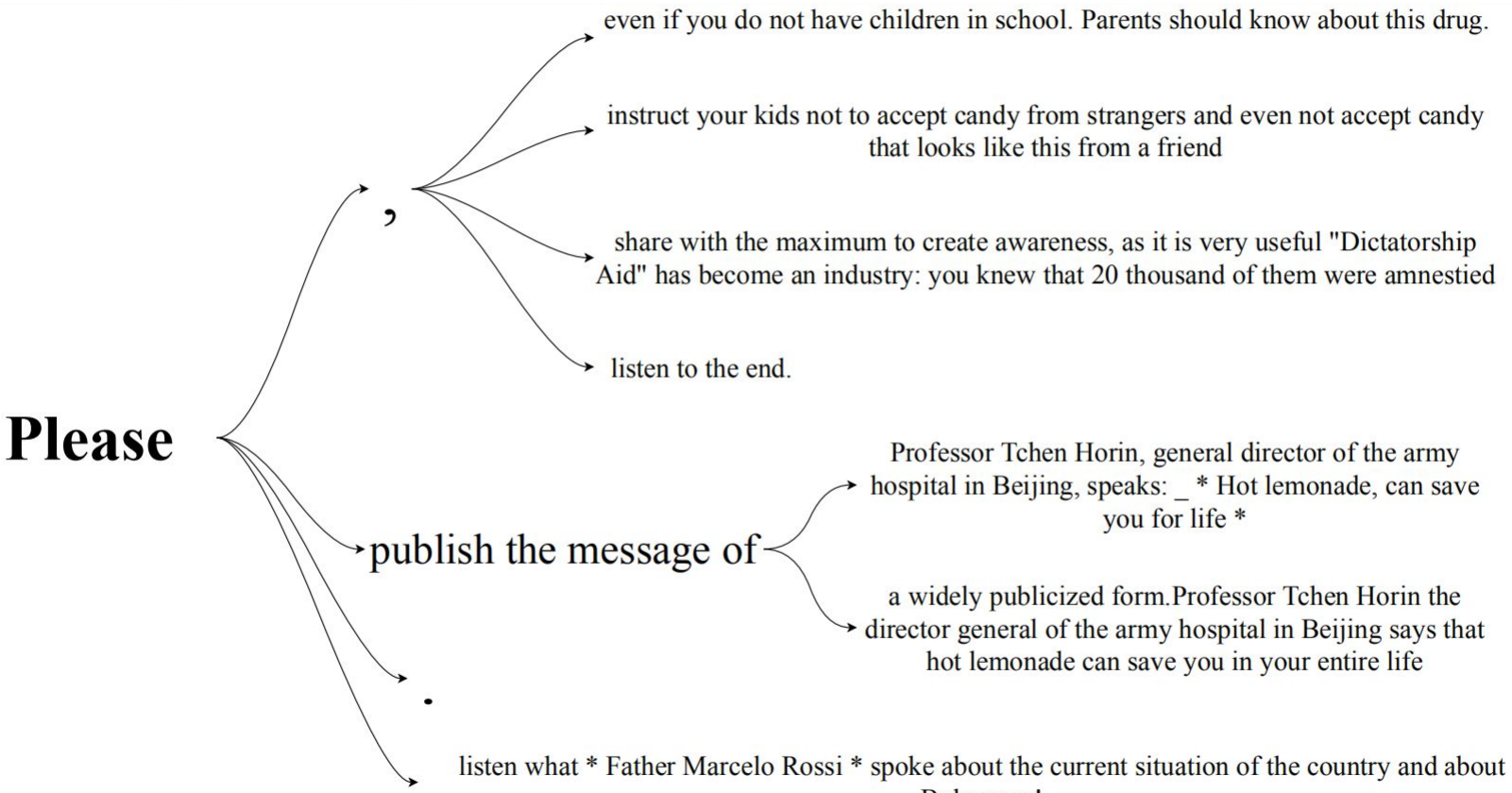
(a) Messages with misinformation



(b) Messages with unchecked content

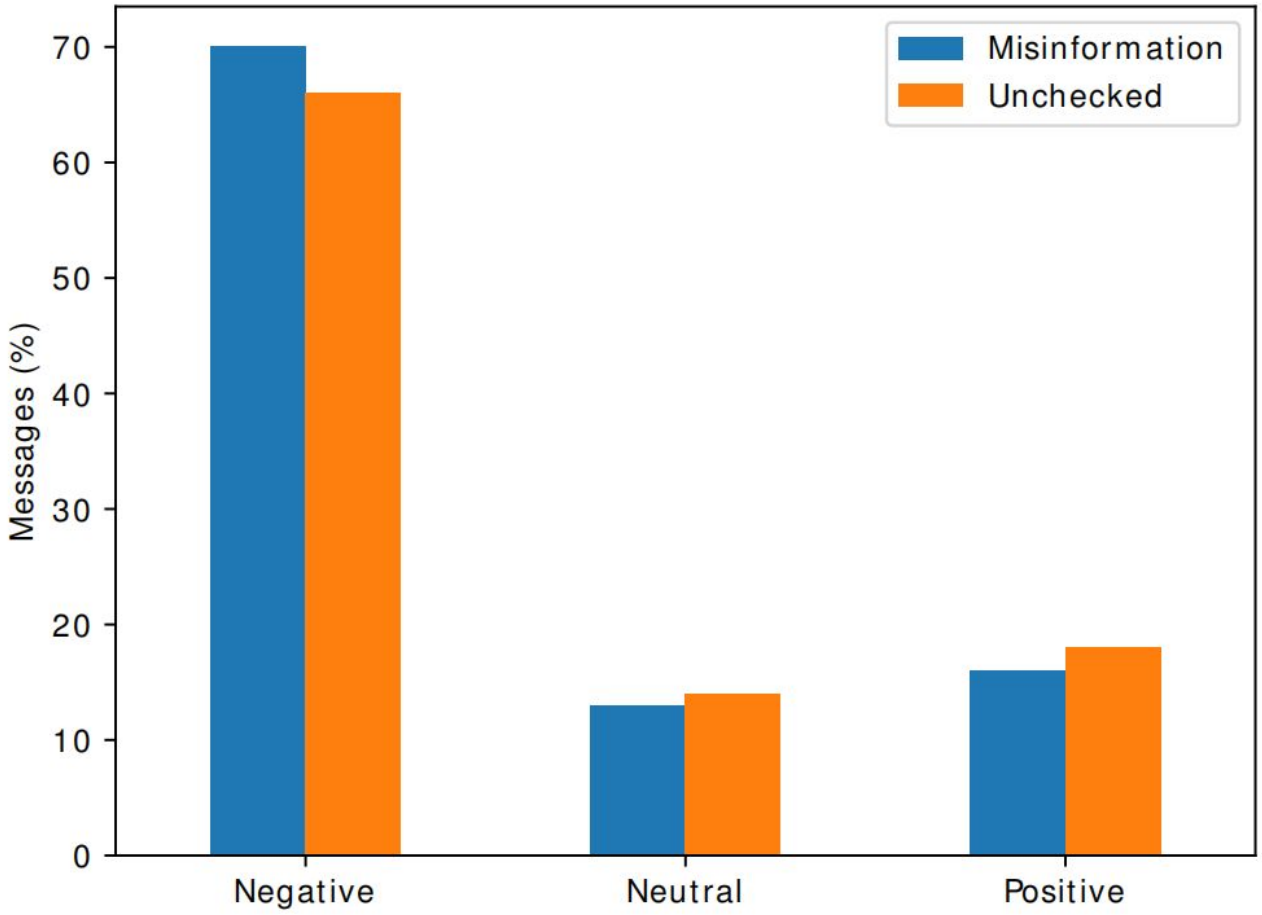
WHATSAPP WEBSCI

Word tree for the word root *Please*



WHATSAPP WEBSCI

Sentiment polarity of messages.



Backup Slides

Data and code

- Here the Colab: https://colab.research.google.com/drive/18-IclaDdOQ3SEaUaZy-J08rVFMqtpHkj
-
- I set some tasks to complete on this document (We could implement more if you think there is time remaining)
-
- - They will download the scripts for tutorial on a gitHub: https://github.com/Phlop/WhatsApp_Tutorial/
- - They will search on Google for links of public groups on WhatsApp (there is also a script there that use a seed of URLs and collect every invitation link from there)
- - They can find the name of a list of groups (following the Kiran's scripts of "getGroupTitles.py"
- - They can join groups on WhatsApp (They can do it manually in their phones or using also the script "joinWhatsappGroups.py", as it needs the Selenium, it will no be possible to run this script on Colab)
-
-

Data and code

- Introduce how to collect data by 2 manners:

Through a crawler in Whatsapp Web using WebWhatsApi: https://github.com/mukulhase/WebWhatsapp-Wrapper (That is how we collect data in Brazil, but they are currently updating the API. I attached a script using this API it use Selenium as well, so needs to configure it too)

Downloading WhatsApp database from the phone:

-The logistic to root the phone and get the key could be hard, so I just created a folder structure online of the phone, so they can interact with the folder of the Android device and find the files we want (key and msgstore.db.crypt12)

They can access this and go through the folders looking for the file http://www.whatsapp-monitor.dcc.ufmg.br/Android/index.html

-After that I left a real example of these files on a folder, so they can run the script "python decrypt12.py key msgstore.db.crypt12 msgstore.db" to see how it works

tools to reconstruct whatsapp group chats

script to vizualize same message over time

script to vizualize images over various platforms