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](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/](https://github.com/Phlop/WhatsApp_Tutorial/)

**Colab:**  [https://colab.research.google.com/drive/18-IcIaDdOQ3SEaUaZy-J08rVFMcqtpHkj](https://colab.research.google.com/drive/18-IcIaDdOQ3SEaUaZy-J08rVFMcqtpHkj) (create a copy and play with the scripts)

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

Top App by Total Minutes

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

- Saudi Arabia: 73%
- Malaysia: 68%
- Germany: 65%
- Brazil: 56%
- Mexico: 56%
- Turkey: 50%
- South Africa: 49%
- Indonesia: 40%
- India: 28%
- France: 23%
- Australia: 22%
- Canada: 19%

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 %)

- Poor: 8% (14%)
- Lower income class: 6% (24%)
- Middle class: 15% (34%)
- Upper middle class and rich: 29% (45%)

Source: Lokniti-CSDS MOTN surveys
Chart 3

The youth are the biggest constituency for WhatsApp

Share of active WhatsApp users (in %)

- 18-25 years: 30%
- 26-35 years: 21%
- 36-55 years: 10%
- 56 years and above: 3%

Source: Lokniti-CSDS MOTN surveys
This is starting to get wide media attention
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
How WhatsApp is fuelling a 'sharing revolution' in Sudan

*Khalid Albaih*

Amid a government crackdown on the press, the app is helping politically charged residents share news and views without fear.
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.
Smartphones, Facebook, WhatsApp break down marriages in Bangalore
Five Killed in Latest Mob Attack After Rumors on Social Media. Here’s What to Know About India's WhatsApp Murders
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.

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

Group size (number of countries)
Users from different countries in a group

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 Many engineering 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](http://whopostedwhat.com) (facebook search)
    - Twitter/reddit search
    - [https://joinwhatsappgroup.com/](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
<table>
<thead>
<tr>
<th>Application</th>
<th>Security</th>
<th>Performance</th>
<th>Audits</th>
<th>Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>2148164711910221;2148164801910212;1548094012:5096029501792714072:5:0:34230</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```
<s class="n_v8i43hbwm t_v8i438vfy b_v8i43gtkt" data-ft="{"tn":"j"}"
    data-testid="story-subtitle">
  ▼<div class="t_v8i43c5_y s_v8i43c603">
    ▼<div class="u_v8i43c5_- s_v8i43c603">
      ▼<div class="y_v8i43c5_- s_v8i43c603">
        ▼<a class="t_v8i439-j8 j_v8i43cvie m_v8i43wab" href="#" role="link" id="u_ps_0_0_0_1j">
          ▼<div class="k_v8i43gpel">
            <span class="s_v8i43c603 z_v8i43f19b">Sp</span> == 0
            <div class="s_v8i43c603 u_v8i43c5_-">Sp</div>
            <div class="s_v8i43c603 e_v8i43anp0">S</div>
            <span class="s_v8i43c603 z_v8i43f19b">on</span>
            <div class="s_v8i43c603 u_v8i43c5_-">on</div>
            <div class="s_v8i43c603 e_v8i43anp0">S</div>
            <span class="s_v8i43c603 z_v8i43f19b">so</span>
            <div class="s_v8i43c603 u_v8i43c5_-">so</div>
            <div class="s_v8i43c603 e_v8i43anp0">S</div>
            <span class="s_v8i43c603 z_v8i43f19b">red</span>
            <div class="s_v8i43c603 u_v8i43c5_-">red</div>
          ▲<div class="k_v8i43gpel">
```

```
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
- 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?

<table>
<thead>
<tr>
<th>Row #</th>
<th>Content</th>
<th>Directory</th>
<th>File</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>contacts database</td>
<td>/data/data/com.whatsapp/databases</td>
<td>wa.db (SQLite v.3)</td>
</tr>
<tr>
<td>2</td>
<td>chat database</td>
<td>/data/data/com.whatsapp/databases</td>
<td>msgstore.db (SQLite v.3)</td>
</tr>
<tr>
<td>3</td>
<td>backups of the chat database</td>
<td>/mnt/sdcard/Whatsapp/Databases</td>
<td>msgstore.db.crypt msgstore-&lt;date&gt;.crypt</td>
</tr>
<tr>
<td>4</td>
<td>avatars of contacts</td>
<td>/data/data/com.whatsapp/files/Avatars</td>
<td>UID.j, where UID is the identifier of the contact</td>
</tr>
<tr>
<td>5</td>
<td>copies of contacts avatars</td>
<td>/mnt/sdcard/WhatsApp/ProfilePictures</td>
<td>UID.j, where UID is the identifier of the contact</td>
</tr>
<tr>
<td>6</td>
<td>log files</td>
<td>/data/data/com.whatsapp/files/Logs</td>
<td>whatsapp.log, whatsapp-&lt;date&gt;.log</td>
</tr>
<tr>
<td>7</td>
<td>received files</td>
<td>/mnt/sdcard/Whatsapp/Media</td>
<td>various files</td>
</tr>
<tr>
<td>8</td>
<td>sent files</td>
<td>/mnt/sdcard/Whatsapp/Media/Sent</td>
<td>various files</td>
</tr>
<tr>
<td>9</td>
<td>user settings and preferences</td>
<td>/data/data/comm.whatsapp/files</td>
<td>various files</td>
</tr>
</tbody>
</table>
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
What’s in this database?

Use e.g. DB browser for SQLite to view the database.
Forensic Analysis of WhatsApp Messenger on Android Smartphones

Cosimo Anglano
DiSIT - Computer Science Institute,
Università 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

<table>
<thead>
<tr>
<th>Field name</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>_id</td>
<td>record sequence number</td>
</tr>
<tr>
<td>key_remote_jid</td>
<td>WhatsApp ID of the communication partner</td>
</tr>
<tr>
<td>key_id</td>
<td>unique message identifier</td>
</tr>
<tr>
<td>key_from_me</td>
<td>message direction: '0' = incoming, '1' = outgoing</td>
</tr>
<tr>
<td>status</td>
<td>message status: '0' = received, '4' = waiting on the central server, '5' = received by the destination, '6' = control message</td>
</tr>
<tr>
<td>timestamp</td>
<td>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)</td>
</tr>
<tr>
<td>received_timestamp</td>
<td>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</td>
</tr>
<tr>
<td>receipt_server_timestamp</td>
<td>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</td>
</tr>
<tr>
<td>receipt_device_timestamp</td>
<td>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</td>
</tr>
<tr>
<td>send_timestamp</td>
<td>unused (always set to '-1')</td>
</tr>
<tr>
<td>needs_push</td>
<td>'2' if broadcast message, '0' otherwise</td>
</tr>
<tr>
<td>recipient_count</td>
<td>number of recipients (broadcast message)</td>
</tr>
<tr>
<td>remote_resource</td>
<td>ID of the sender (only for group chat messages)</td>
</tr>
</tbody>
</table>

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](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
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
Opinion

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

By Cristina Tardáguila, Fabricio 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
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
Political groups

In India, Facebook’s WhatsApp Plays Central Role in Elections
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.

Janta Ki Baat survey predicts huge BJP gain in Karnataka assembly elections.

★ BJP - 135 Seats
◆ JDS - 45 Seats.
● Congress - 35 Seats
● Others - 19 Seats

The latest poll survey, conducted by Janta Ki Baat with the sample size of a 10.20 lakh respondents, says BJP is crossing 135 seats in Karnataka assembly elections.
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

WHATSAPP MONITOR

Brazil
WhatsApp Monitor for Brazil

Access»

India
WhatsApp Monitor for India

Access »

Indonesia
WhatsApp Monitor for Indonesia

Access »
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/

Disinformation Spreads on WhatsApp Ahead of Brazilian Election

# Table 1: Overview of our datasets.

<table>
<thead>
<tr>
<th></th>
<th>Truck Drivers’ Strike</th>
<th>Election Campaign</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Groups</td>
<td>141</td>
<td>364</td>
</tr>
<tr>
<td>#Total Users</td>
<td>5,272</td>
<td>18,725</td>
</tr>
<tr>
<td>#Total Messages</td>
<td>121,781</td>
<td>789,914</td>
</tr>
<tr>
<td>#Text Messages</td>
<td>95,424</td>
<td>591,162</td>
</tr>
<tr>
<td>#Images</td>
<td>11,610</td>
<td>110,954</td>
</tr>
<tr>
<td>#Videos</td>
<td>9,752</td>
<td>73,310</td>
</tr>
<tr>
<td>#Audios</td>
<td>4,995</td>
<td>14,488</td>
</tr>
<tr>
<td>#URLs</td>
<td>11,728</td>
<td>92,654</td>
</tr>
</tbody>
</table>
Events Captured in Brazil

![Graph 1: Truck Driver's Strike](image1)

**May 21\textsuperscript{st} - Jun 2\textsuperscript{nd} 2018**

- Images
- Audios
- Videos

![Graph 2: Election Campaign](image2)

**Aug 16\textsuperscript{st} - Sep 7\textsuperscript{th} 2018**

- Images
- Audios
- Videos

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

- WhatsApp data
- Gather Images from WhatsApp
- Search for images on Google
- Web pages returned by search

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.

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.
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.
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

\[ \Delta T = T_{\text{last}} - T_{\text{first}} \]

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

<table>
<thead>
<tr>
<th>Election Campaign</th>
<th>Misinformation</th>
<th>Unchecked</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Groups in which images were shared</td>
<td>157</td>
<td>351</td>
</tr>
<tr>
<td>#Users who shared images</td>
<td>624</td>
<td>10,339</td>
</tr>
<tr>
<td>#Unique images</td>
<td>85</td>
<td>69,590</td>
</tr>
<tr>
<td>#Total shares of images</td>
<td>1,168</td>
<td>109,791</td>
</tr>
</tbody>
</table>

- 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!

Brazil
WhatsApp Monitor for Brazil

India
WhatsApp Monitor for India

Indonesia
WhatsApp Monitor for Indonesia

Your country here!
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.
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

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 |

https://www.independent.co.uk/news/world/europe/spain-elections-whatsapp-podemos-channel-close-left-inq-de-olmo-a8886481.html
Case 2: Brazil university budget cuts
Large scale micro targeting!

Images claiming to be from universities started spreading on WhatsApp groups.
Large scale micro targeting!

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

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.

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

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

"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.

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

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 international 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

Mangusingh Rajpurohit @mangusinghBjp4 · Jan 9
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

Chowkidar रवि गुप्ता @RaviVym · Jan 9
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

Dr. Bholasinh G Patel @MdDrBholasinh · Jan 9
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

Pankaj Vyas @erpankajvyas · Jan 9
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

- 4,645 Webpages
- 5,208 Webpages
- 425.2 Days
- 5.8 Days
- 1,168 Shares

WhatsApp

Images with unchecked content

- 78,232 Webpages
- 36,599 Webpages
- 510.9 Days
- 9.4 Days
- 20,377 Shares

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**

**Facebook 2019-04-21 13:26:04**

**Sharechat 2019-04-19 09:26:10**
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.
The way of reclining

Ok I am married and I have 4 children about

I mean 6 people, my wife and 4 children

How much is the cost per one

What brother

Each win is 1400 euros

Continue to Thessaloniki

I mean how much is in dollars

1600 $ 

I mean 16 sheets

Right

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
2. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C2&q=wa.db+whatsapp&btnG= (list of related work analyzing whatsapp database)
5. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures (WWW 2019)
7. Forensic analysis of encrypted instant messaging applications on Android
Thank you!
Additional slides
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
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

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

(a) Messages with misinformation

(b) Messages with unchecked content
Word tree for the word root *Please*

- Even if you do not have children in school, parents should know about this drug.
- Instruct your kids not to accept candy from strangers and even not accept candy that looks like this from a friend.
- Share with the maximum to create awareness, as it is very useful. "Dictatorship Aid" has become an industry: you knew that 20 thousand of them were amnestied.
- Listen to the end.

*Please*

- Publish the message of...
  - Professor Tchen Horin, general director of the army hospital in Beijing, speaks: *Hot lemonade, can save you for life*.
    - A widely publicized form. Professor Tchen Horin, the director general of the army hospital in Beijing says that hot lemonade can save you in your entire life.
    - Listen what *Father Marcelo Rossi* spoke about the current situation of the country and about Bolsonaro!
WHATSAPP WEBSCI

Sentiment polarity of messages.

![Bar chart showing sentiment polarity](chart.png)
Data and code

• Here the Colab:
  https://colab.research.google.com/drive/18-lclADdQQ3SEaUalZy-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