Engagement index for users and conversations in encrypted messages from WhatsApp groups

Moshe Cotacallapa*, Didier A. Vega-Oliveros†‡

*National Institute for Space Research (INPE), Brazil / moshe@usp.br
†Department of Computing and Mathematics - University of São Paulo, Ribeirão Preto, SP - Brazil / davo@icmc.usp.br
‡School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN - USA

Abstract—WhatsApp, a very popular cross-platform messaging platform with more than one and a half billion users, is the preferred medium for communication in several countries. One of the interesting features in the platform is called Groups (WG), which is an encrypted end-to-end virtual room created by a group of individuals, where only the group members can send and see the messages. The goodness of privacy and security offered by WG have attracted the attention of families, friends, businesses, and organizations. However, in several cases, those WG’s have been used as a powerful tool for spreading dangerous content or fake news. In this work, we propose a new method for measuring the level of engagement in conversations and users from WhatsApp groups, by using temporal interaction networks and without reading the messages. Our framework creates an ensemble of networks that represent the temporal evolution of the conversation every 10 minutes. In this way, we use network measurements to build an Engagement Index (EI) for fractions of the conversations. Our results in five real-world WGs data indicate that the EI is able to identify different type of conversations, and users’ behaviors according to each category, as well as anomalies on users’ engagement.

Index Terms—User characterization, Network analysis, Temporal Networks, Encrypted group messages, Engagement index

I. INTRODUCTION

Online Social Networks have become the new channel for sharing and spreading information. The diffusion of rumors, innovations, knowledge, and news are very common in social networks like Twitter, Instagram, Sina Micro-blog, or Facebook. In particular, citizens from several countries have adopted Mobile Instant Messaging apps, like WhatsApp, Telegram, etc., as the preferred medium for communication with their family, friends, coworkers, or clients. Moreover, many of these platforms provide the possibility of group communication, which allows massive and viral communication flows between group users for sharing information rapidly. For example, WhatsApp Groups (WGs) were one of the main arenas for intense political or marketing campaigns, self-organized movements, among other activities in many countries. In India, for instance, the two major political parties claimed to have more than 20 thousand WGs that allowed to mobilize millions of sympathizers [1]. The parties also had thousands of “WhatsApp warriors” to participate in groups posting biased, inflammatory political content or fake news. In Brazil, an audio with fake information about a new flu pandemic with mortal victims was shared among several WGs, producing collective paranoia and chaos in public health services [2].

Due to privacy issues in many popular social and messaging platforms, a huge quantity of private conversations and personal information from their users were used for monitoring citizens, political campaigns, targeted ads, and many other initiatives from third-party companies and governments. In order to improve the security and privacy in the communication, many of the Mobile Instant Messaging apps have implemented end-to-end encrypted communication between users. However, the goodness of privacy and security offered by the system can be used for illegal activities or problematic behaviors, like the spread of rumors, spams, fake news, and the influence of the public opinion for any goals.

Here, we aim to develop a framework for analyzing and classifying the users and conversations in encrypted message groups. Based on the construction of temporal interaction networks of the conversation, we introduce the Engagement Index (EI), an alternative approach for measuring the level of engagement of conversations in encrypted groups. Experimental results shown that the proposed index is capable of measuring the engagement of individuals over time. Furthermore, the temporal snapshots of the conversation can be classified in categories according to z-score dispersion of the EI values, and therefore, mining which users are more involved on each conversation category. Moreover, it was possible to find anomalies on users’ behavior during the Brazilian president electoral period by employing the time-series of EI centrality. Our main contributions are threefold: 1) A flexible framework for measuring the engagement of conversations and users, without considering the content of messages; 2) A classification of chats and ranking of users according to their engagement values; 3) The novel EI for analyzing and mining users’ behavior, detecting, for instance, anomalous users.

II. RELATED WORK

It is increasingly accepted that many natural and social systems are better described as temporal or dynamic networks [3], with links that exist only at certain times. Some examples are: the temporal and dynamical user’s behavior in social networks [4]; the temporal evolution of disciplines guided by scientists social interaction [5]; calculating the analytical epidemic threshold [6] or controlling the spreading
of epidemics or information \[7\], \[8\]; and, understanding the spatiotemporal evolution of wildfires according to the community identification \[9\].

Several works have developed methods to measure the engagement of individuals in social environments, like the discovering of correlations between the use of Facebook and student engagement \[10\]; the role of Twitter opinion leadership on individuals’ engagement in politics \[11\], and measuring the engagement of businesses on Twitter and the positive effects on consumers’ engagement with online word-of-mouth communication \[12\]. Another approach, based on content analysis, is focused on studies applying Natural Language Processing (NLP) techniques for processing User Generated Content in Opinion Mining and Sentiment Analysis \[13\]. Also, based on these techniques, several studies have developed methods to measure the users’ engagement on marketing campaigns, products or individuals, over time \[14\], \[15\]. However, these previous work employed the content of the messages as input of the methods and are not suitable for the case of encrypted scenario. To the best of our knowledge, this is the first approach that tackle the mining of users’ interaction and engagement in encrypted group message and by employing temporal networks.

III. RESEARCH PROBLEM

Instant Messaging apps have adopted features that allow users to communicate in groups. Moreover, the platforms provide encrypted end-to-end communication where only group members can read and share the messages. However, the goodness of privacy and security make these groups a powerful tool for influencing public opinion about politics, brands, spreading malicious content or fake news. For that reason, approaches based on Content Analysis are not suitable to measure the users’ engagement in this scenario, since it is not possible to gather messages content. Besides, it is not possible to access complementary information, like followers and following users, or likes (in the case of Twitter), comments and profile information (in the case of Facebook).

Due to privacy concerns, several social and messaging platforms are deploying tools to encrypt users’ messages. Considering that in WhatsApp groups (WG) we do not have access to any personal information or messages since they are encrypted, the critical point to tackled here is: How to measure users’ engagement on WG without reading the content of their messages? We break down this point into the following research questions:

- How to appropriately represent and construct the network of user interaction over time? Moreover, how to characterize the networks?
- Can we classify users according to the pattern of interaction and establish their levels of engagement within the group?
- From the behavioral pattern of users, could be possible to identify the following profiles?: (1) The most engaged or (2) influential users; and (3) similar patterns of behavior?

This work is an endeavor in this direction, where we aim to analyze the users’ behavior when sending messages in encrypted mobile/online apps. Our approach consists in construct an Interaction Network from the messages sends by the users in the specific case of some WG.

The construction is related to the messages sent in the group in a time interval (\(\Delta t\)). Therefore, we obtain an ensemble of interaction networks that represent the temporal evolution of the user’s activity into the WG. A natural path for analyzing this ensemble is to employ temporal networks \[3\], \[9\], \[16\], where each layer represents a snapshot of user’s activity in a particular \(\Delta t\), and the intra-layer connections represent the temporal evolution of users. Therefore, this temporal approach allows the mapping of behavioral patterns of social interaction in local, intermediate and global scales of the evolutionary process.

IV. MATERIAL AND METHODS

In network sciences, the broadest approach is to mathematically represent data by a static graph \[1\] \(G = (V,E,W)\), where \(V = \{v_1,v_2,\ldots,v_n\}\) is the set of \(n\) agents called as nodes, the set of \(m\) edges or links \(E = \{e_1,e_2,\ldots,e_m\}\) that connects the nodes and the set \(W\) of \(m\) weights, one for each edge. However, when consolidating the temporal information in static representation, we lost part of the network evolution. In this sense, it is not possible to evaluate the performance and role of the nodes, nor understanding the interaction patterns into the network.

As alternative, the dynamical or temporal network \(G)\) can be represented as an ordered sequence of network observations at different time-steps or intervals \[16\], i.e., \(G = \{G_0,G_1,\ldots,G_l\}\) with \(l\) the number of layers or snapshots \[3\], \[16\]. In other words, for temporal networks, we have a long sequence of symmetric pairwise interactions representing observations over time. This dynamical network contains not only the set of similarity or relation links between nodes but also information on how the connection behavior evolves.

Nevertheless, many times the data is not “naturally” represented by a graph, but rather by events or time series. Thus, for applying the network techniques, a network must be constructed. One approach for reconstructing data events into networks is upon the process of linking nodes according to the co-occurrence of events in a chronological fashion \[9\]. In Earth Sciences, \[17\] employed an approach called sequential networks, for describing and finding patterns on the earthquake network of the United States of America and Japan. \[18\] model the moisture recycling process in South America and developed a similar framework to event-based networks with weighted nodes and directed links. In the same line, \[19\] obtained a better understanding of long-range seismic activities by introducing window times and new rules to link nodes. Recently, \[9\] analyzed spatiotemporal wildfire events into chronological networks.

\[1\]The terms “network” and “graph” share the same definition and are interchangeable in this document.
Finally, the engagement characterization is performed based on different analysis for each interaction network. The codes and data produced during the development of this project will be made publicly available on the Internet. In the next section, we present the network construction and the proposed framework for measuring the Engagement of users’ interaction.

B. Modeling messages behavior via interaction networks

The users’ patterns of interactions play a fundamental role to define their ability to propagate information or influence in their group. However, we need first to reconstruct the data into a network of interaction. Thus, with these networks, we can characterize the dynamical and topological properties of the groups related to the spreading ability of users, problematic behaviors, or engagement in the conversation. By identifying these features, we can classify the users’ behavior by looking on the interaction patterns, i.e., we can understand the group dynamics when sending information and predict the users’ influence based on their activity behavior, instead of taking into account the content of the messages.

We analyze the users’ behavior when sending messages in WGs by employing temporal networks. Our network approach consists of constructing Interaction Networks from the messages sent by the users in the WG. The construction is related to the messages sent in the group in a specific time interval ($\Delta t$). Therefore, we obtain an ensemble of interaction networks that represent the temporal evolution of the user’s activity into the WG. This modeling allows the mapping of behavioral patterns of social interaction in local, intermediate and global scales as an evolutionary process.

The main idea of our framework is illustrated in Figure 2. We select as an example, the message activity from Politics1 WG collected over three days. Figure 2.A shows the count of messages in intervals of $\Delta t = 10$ minutes. As we collected those messages in plain text, for illustrative purposes, we tagged some moments with the particular topic discussed at the moment. However, it is important to highlight that we do not use the content messages on our method.

Also, to illustrate how a discussion is formed, in (Figure 2.B) we show the four stages of a discussion. In the first moment, there was no discussion or significant message activity. Suddenly, the beginning stage of a discussion is trigger by some initiator users. After that, the discussion reached a broad message activity and the climax with increased participation, in the middle stage. Finally, the discussion diminishes and gets to the resolution finishing the stage.

The process to transform the message activity into an interaction network is by considering the co-occurrence of messages sent by users. In Figure 2.C, we show the network construction process for a hypothetical conversation. Node A corresponds to Phone A, node D to Phone D and so on. Each time a node sends a message, it is connected to the node of the previous message. For instance, node D sent a message after A at Time2, and thus, in the interaction network, they are connected. We generate weighted and undirected networks avoiding self-loop connections. Here, the multi-edges are...
represented as the sum of edges between two nodes and used as the weights of the connections. This graph representation, which is a snapshot of the message activity of users, differs from only considering the number of messages. For example, if a user sends 30 messages in a row, s/he is merely interacting alone, with a low group engagement in the conversation. For this reason, we define that is necessary at least two users interacting to be considered as a conversation.

As shown before in Figure 2B, a discussion has different stages over time. Then, by using the network characterization for the same discussion, we generate a sequence of networks, each one representing time slices of interactions. The generated networks present several structures, indicating that the interaction networks have no trivial or regular connections. In this way, we can extract some interaction patterns from the topological information of the ensemble. For instance, the network at 23:30 of July 15 has 4 nodes as the midnight network of the same day. However, the structures are different. For this reason, we present a new measure called Engagement index, which seeks to quantify the engagement regarding users’ interaction on the network, presented as follow.

C. The Engagement index (EI)

The EI of a network quantifies how is the level of interaction between the nodes, i.e., how are distributed the interactions (links) among the user’s participation in terms of the network. We derive the EI reflecting the number of interactions between nodes, taking into consideration how equally distributed were the participation. For this purpose, we evaluate not only the quantity of messages, but how many users were interacting and how equally was that interaction. Formally, we present the definition of the EI in terms of the equality and intensity interaction of the network, as follow:

\[
\text{Engagement}(G) : EI(G) = \text{Equality}(G) \times \text{Intensity}(G) \\
\text{Equality}(G) = 1 - Gini(W) \\
\text{Intensity}(G) = \log_2(n \times \frac{1}{2} \sum_i w_i)
\]

where the Engagement is the product between the Intensity and Equality of users’ interaction on the network.

The Equality is the complement of the Gini coefficient, originally proposed for measuring the level of inequality in the incoming of a population [20]. The Gini values vary from 0 (full equality) to 1 (total inequality). Therefore, in Eq. 2 we are interested in measuring how equally was the message interactions (weighted links) among the participants.

In Eq. 3 we have that all \( w_i \in W \) are positive integers greater than zero. With the Intensity we measure how intense was the conversation (network) in terms of the number of participants (nodes) and the total user-to-user messages (links). The Intensity is equal to 1 when the network has at least two nodes interacting once, i.e., with an average degree equal to 1. The before is the reason for the \( \log_2 \) in the measure.
TABLE I
TOY NETWORKS DEPICTING DIFFERENT CASES OF APPLYING THE PROPOSED ENGAGEMENT INDEX. FOR ILLUSTRATIVE PURPOSE, WE PLOT THE NETWORKS SHOWING MULTI-EDGES INTERACTIONS.

\[
\begin{array}{cccc}
\text{Index} & \text{(a)} & \text{(b)} & \text{(c)} \\
\text{n} & 4 & 3 & 3 & 2 \\
\text{m} & 8 & 6 & 6 & 4 \\
\text{Intensity} & 5 & 4.17 & 4.17 & 3 \\
\text{Equality} & 1 & 1 & 0.83 & 1 \\
\text{Engagement} & 5 & 4.17 & 3.47 & 3 \\
\end{array}
\]

The case of a network with a single node, it is not considered a group conversation and the Intensity is equal to zero.

The EI is then the combination of high message activity between a large number of individuals but with a more equally distributed participation. For example, a conference talk or broadcasting communication, where only one source is expressing its ideas, is not ideally considered as a conversation with high engagement. In Table I, we present some illustrative examples of interaction networks and their EI values.

First, we have in Table I columns (a), (b), and (d) three network with homogeneous distribution of links between the nodes. Therefore, the Equality of the networks is the highest. However, the network in (a) reach a higher EI due to a larger number of participants than networks (b) and (d). On the other hand, the network in (c) has the same number of nodes, links, and Intensity than network (b) but most of the interactions are concentrated in only two nodes, producing a drop in the Equality of the network (c). For this reason, network (c) achieve lower Engagement value than network (b).

Regarding the EI for the nodes, we define the Engagement centrality as the EI value of the network proportional to the participant interaction with respect of all the nodes, i.e.,

\[
EI(G, v_i) = \frac{n \times w_{k_i} \times EI(G)}{2 \times m}
\]

\[
EI(G) = \langle EI(G, v) \rangle
\]

where \( w_{k_i} \) is the weighted degree of node \( v_i \). In this way, each node contributes proportionally to the network Engagement according to their number of interactions (\( w_{k_i} \)), which lead to EI of the network is the average of the EI of the nodes.

V. EXPERIMENTAL RESULTS

We applied our proposed framework to construct the temporal networks of interactions for each WG. The distribution of edges and nodes for the five WG in 60 days of data are shown in Figure 3. We observe that most of the networks have message interactions between 2 and 5 participants (nodes). However, in controversial groups related to the discussion of a political point of view (Figures 3(a) and 3(b)), the number of user-user interactions (edges) is higher than the other groups. The Vegetarian (Figure 3(c)) and English (Figure 3(d)) groups present the lower number of maximum nodes and edges.
In order to compare the similarities and differences across the groups, we calculate the z-score values of the networks by WG, in the way,

$$z\text{-score}(G_j) = \frac{EI(G_j) - \text{MEAN}(EI(G))}{\text{STD}(EI(G))}, \quad \text{with } G_j \in G$$  \hspace{1cm} (6)

where MEAN(EI(G)) means the average of the EI for all the networks in the WG G, and similar for the STD. In Figure 4 we present the histograms of the z-score values for each WG.

We can observe that each WG has particularities in terms of the level of EI over the networks. In Politics2, English, and Theology groups, a portion of the networks with EI have one standard deviation below the mean. Another fraction of networks have EI values close to the mean, and some other networks have high engagement values, above one standard deviation of the mean. Given the before considerations, we classify the networks in three z-score categories (as shown in Table II): HIGH engagement networks, with z-scores values greater than or equal to 1; MEDIUM engagement networks, with z-score values between (−1, 1); and LOW engagement networks, with z-score values below or equal to −1.

| Groups      | Categories | Total Networks |
|-------------|------------|----------------|
| Politics1   | LOWER 17%  | 2879           |
| Politics2   | MEDIUM 67% | 2535           |
| Vegetarian  | MEDIUM 19% | 802            |
| English     | HIGH 29%   | 1091           |
| Theology    | MEDIUM 20% | 2417           |

Table II: Proportion of conversations (networks with n > 1) in each category for the five WGs.

LOWER engagement networks can be identified as initiators, claimers, or finishers of the discussion topics of the group. Opposite, representative users in the HIGHER EI networks, can be seen as conciliators, or argumentative users expressing strong positions in the conversation. Besides, depending on the discussed topic, a prominent engaged participant could be less active or not interested in sharing its opinion. Thus, it is also interesting to consider a global or general score of engagement of users through the discussions in the group.

We extended the network EI classification to the nodes of the networks. First, the EI centrality is calculated for the nodes of all the WG networks. In Figure 5 we show the EI centrality for six users from the Politics2 group. Each bar represents the temporal Engagement evolution of the user over time. We notice that this time-series characterize the message behavior of the users, showing the moments in which the users interact with the others and their levels of engagement or relevant participation in the group.

We separate the networks into three groups according to the EI classification, as shown in Table III. Then, for each node in the group of networks, i.e., in the LOWER, MEDIUM, or HIGHER groups, we calculate the average EI centrality in the group. Additionally, we calculate the average EI centrality of the nodes considering all the networks, which we call as the GLOBAL group. As an example, the top 10 ranking of EI centralities by EI classification are reported for the Politics2 WG in Table III. The IDs of the users were renamed according to the GLOBAL ranking, and are the same IDs used in Figure 5.

In the Table III we have that the GLOBAL top ranked user

Fig. 5. Engagement values for six particular users from Politics2 WhatsApp Group.

The proposed classification of the network is vital for understanding the dynamic of high, medium, or low engagement in group conversations. This network classification also can be extended to the nodes, in which the participants can have different roles depending on the message interaction they have in the network classes. Users that are more representative in
is also the best ranked in the HIGHER and MEDIUM group. In this particular case, this user is the moderator/manager of the group, which is a very active participant. However, this is not the natural tendency for the other groups. Comparing the rankings in each classification group, we can observe there are some position differences between the users. The difference in Engagement behavior is notable according to the ranking in the classes and the Figure [5] ID 5 has more regular participation during the discussion over time, but it is better ranked in MEDIUM and LOW EI networks. Opposite, ID 0 and ID 1 have meaningful participation in (GLOBAL) and in HIGH EI networks. However, user ID 1 tends to be a little engaged in LOW or MEDIUM EI networks. On the other hand, ID 7 and ID 39 are better ranked in LOW EI networks than in the other classification groups, indicating the tendency of interacting in low activity moments.

VI. APPLICATION CASE: USERS ANOMALY BEHAVIOR

We present an application of the proposed framework for identifying anomaly behavior of users over time when it is not possible to read the content. We aim in this study case to identify the differences in the users’ behavior on three WGs: Politics2, Vegetarian and Theology, specifically during the presidential elections that happened in Brazil in 2018. Therefore, the following two intervals where considered:

\((P_1)\) Pre-electoral: From first of October to October 28 of 2018.

\((P_2)\) Post-electoral: After Brazilian presidential elections, from October 29 to November 21 of 2018

To discover whether the presidential election has a different influence over the Politics group than the others, as a first step, we characterize both periods (Pre-electoral and Post-electoral) using the proposed framework of temporal interaction networks. The EI values are calculated for all conversations (networks with \(n > 1\)). Then, we calculate the average EI centrality for each user considering: (i) the whole period \((P_1 + P_2)\); (ii) the separate periods \(P_1\) and \(P_2\) independently. In each case, we obtain a vector of the average EI centrality of the nodes, which are normalized by the highest value of the vector. Next, we sort the users by EI centrality in terms of the whole period.

We compare the differences between \(P_2\) and \(P_1\) for all the users. This way, negative values mean that the EI of the users before the presidential elections were higher than after the elections, and positive values otherwise. The top 100 users in the whole period and their differences values are on Figure [6] Each line corresponds to one of the three WGs. Notably, the difference between \(P_1\) and \(P_2\) is more intense for Politics2 WG than the other two WGs. We observed that some users reached more negative difference values. Five of the top members (within top 10) achieved values lower than \(-0.61\). The before indicates that this set of users, who had a high engagement before the elections, abruptly stop interacting after the first or second round.

To visualize how those five users were engaged over time, in Figure [7] we show their corresponding EI values, where each network represents a conversation. Note that although those users lose engagement, some of them are still active on some points after the second round (point \(\beta\) in the figure). It is interesting to highlight that the proposed index is flexible and suitable for different time windows. For example, we could compare the gain or loss of engagement in weekly or monthly intervals, or even in real-time, and finding who are the members that were engaged and are not anymore.

VII. CONCLUSION

In this work, we proposed a framework for analyzing users’ behavior in encrypted group message applications. We employ temporal interaction network for representing the user-user message interaction over time. Given the encryption constraint, we introduce the Engagement index for measuring the level of participation and engagement of the users according to their interaction behavior. We tested the framework with data collected from five groups of WhatsApp.

By mining this data, it is possible to make a more accurate group description and recommendation, identifying key or influential users by topics, or suspicious behavior in both groups and users scale. In our understanding, this project contributes opening a new path for identifying interaction

| Table III: Ranking of the nodes according to the average Engagement of the networks for the Politics2 WG, according to the classification: the group of HIGHER Z-score values; the group of MEDIUM Z-score; networks in the group with LOWER Z-score values; and the GLOBAL or overall ranking among all the valid networks. |
| --- |
| **HIGH** | **MEDIUM** | **LOW** | **GLOBAL** |
| ID | Mean | ID | Mean | ID | Mean | ID | Mean |
| 0 | 2.719441 | 0 | 0.787287 | 0 | 0.112828 | 0 | 0.954576 |
| 1 | 2.100645 | 2 | 0.619459 | 5 | 0.111283 | 1 | 0.686651 |
| 2 | 1.611634 | 5 | 0.559691 | 3 | 0.092736 | 2 | 0.659132 |
| 4 | 1.540280 | 3 | 0.555109 | 6 | 0.069552 | 3 | 0.570785 |
| 3 | 1.283914 | 1 | 0.522038 | 2 | 0.061824 | 4 | 0.554785 |
| 6 | 1.137957 | 7 | 0.519093 | 39 | 0.058733 | 5 | 0.491393 |
| 8 | 1.082237 | 4 | 0.465320 | 0 | 0.054096 | 6 | 0.477436 |
| 5 | 0.811269 | 6 | 0.447048 | 20 | 0.051005 | 7 | 0.410872 |
| 10 | 0.783183 | 9 | 0.265868 | 26 | 0.044822 | 8 | 0.343968 |
| 11 | 0.725720 | 8 | 0.249979 | 50 | 0.043277 | 9 | 0.269298 |

...
users’ patterns in encrypted messages groups. For example, the presence of annoying users, such as spammers, bots or, fake profiles that promote brands or political parties, can produce a terrible experience for the rest of the members and the propagation of fake news or hoaxes.

As future works, several analyzes can be performed following or extending the proposed framework, applying not only in WGs but also in other similar environments, like Telegram Groups, Forums or Live stream chats. Besides, with the proposed framework could be possible to detect temporal patterns of interaction, calculate the engagement correlation between users, and the characterization of group topics according to the users’ interaction over time. Furthermore, the interaction networks can be analyzed considering motif-based patterns [21], like in ecological or food web networks, where it could be characterized WGs with similar topics by the corresponding motifs counts.

ACKNOWLEDGMENTS

The authors thanks the Sao Paulo Research Foundation (FAPESP) (Grant 2015/50122-0) and DFG-GRTK (Grant 1740/2) for the financial support. DAVO acknowledges FAPESP (Grant 2018/24260-5, 2016/23698-1, and 2018/01722-3).