Community-Driven Fact-Checking on WhatsApp: Who Fact-Checks Whom, Why, and With What Effect?

Kiran Garimella

kiran.garimella@rutgers.edu

Abstract

This paper studies community-driven fact-checking –the members of a community fact-checking their own content– on WhatsApp, with the aim of determining its prevalence, who does it, and whether it is effective. The study leverages two large datasets of WhatsApp group chats, encompassing both public and private group conversations with varying levels of intimacy among members. Adopting a mixed-methods approach, the research combines quantitative analysis of observational data with qualitative measures to shed light on these research questions.

The findings reveal that community-driven corrections are infrequent, and when they do occur, they are typically conveyed through polite requests aimed at alerting individuals about the presence of misinformation. However, users often exhibit apathy towards self-correction, disregard the corrections, or even feel offended by public corrections within the group. Notably, the responsibility of correcting misinformation primarily falls on active community members, with group administrators accounting for a relatively small portion (up to 20%) of the corrections. Additionally, the study uncovers significant variations in the types of corrections and responses to corrections, influenced by group norms and the degree of familiarity among group members. These observations suggest the existence of underlying dynamics of power and trust within these groups.

The insights from this research hold implications for fact-checking and policies in encrypted chat platforms, as well as the role of the community in facilitating accurate information dissemination even in non-encrypted discussions. By shedding light on the efficacy and contextual factors surrounding community-driven fact-checking, this study contributes to a deeper understanding of fact-checking practices on encrypted chat platforms, paving the way for informed interventions and strategies in the realm of misinformation mitigation.

1 Introduction

The proliferation of misinformation on online platforms has raised concerns about its potential impact on society. While extensive research has been conducted on fact-checking practices ([Porter and Wood2021]), the majority of studies have focused on non-encrypted platforms, neglecting the distinctive challenges posed by encrypted platforms like WhatsApp. Because even the platforms can not see the content spreading on them, these encrypted platforms present unique difficulties for traditional top-down, reactive fact-checking approaches, necessitating alternative strategies to combat misinformation effectively.

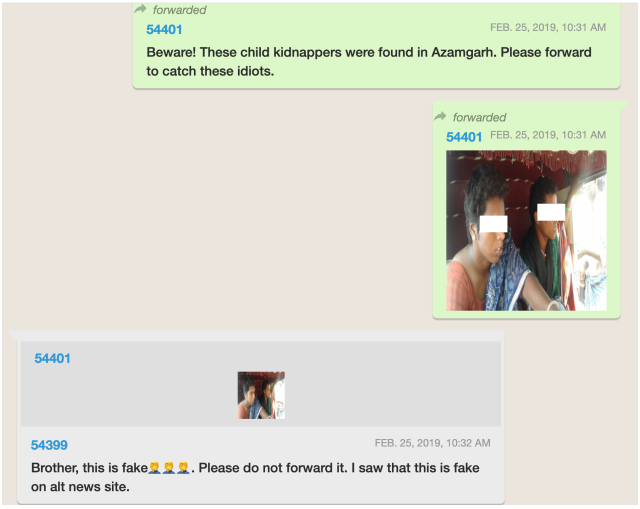


Figure 1: An instance of in-group, community-driven fact-checking. A user (anonymized ID 54401) forwarded an image of alleged child kidnappers, and another user (ID 54399) replies debunking it. Such rumors of child kidnappers have killed dozens of people in India. In most cases, such information goes unchecked, and hence can become life threatening ([Arun2019]).

Community-driven fact-checking has emerged as a promising approach to address the limitations of traditional fact-checking methods, particularly within the context of end-to-end encryption. Community-driven fact-checking (see Figure 1) refers to an instance of correcting misinformation by members within a community/group ([Kligler-Vilenchik2022]). This approach harnesses the collective power of the community to identify and correct misinformation, leveraging the diverse knowledge and resources of group members. However, empirical studies investigating the effectiveness of community-driven fact-checking on WhatsApp remain scarce.

This paper aims to bridge this gap by examining politi-

cal conversations on WhatsApp to explore the prevalence, dynamics, and efficacy of community-driven fact-checking. To achieve this objective, we use two large-scale datasets of WhatsApp group chats, encompassing both public and private conversations with varying levels of intimacy among group members. We adopt a mixed-methods approach, combining quantitative analysis of observational data with qualitative measures to provide comprehensive insights into community-driven fact-checking practices.

Our findings highlight several key observations. Firstly, community-driven corrections are infrequent, indicating a potential reluctance among users to engage in a correction in the public. Most attempts at corrections are polite requests aiming to inform individuals about the presence of misinformation without causing offense. The responsibility for social corrections is primarily shouldered by active members of the community, with group administrators contributing to only a small portion of the corrections (up to 20%). The reactions of users to these corrections vary, with some displaying indifference, others disregarding the corrections, and some feeling offended by public corrections within the group. Finally, our analysis reveals that there are no short or long term effects to being subject to a correction.

By analyzing multiple datasets with diverse social contexts (public groups with members who may be strangers, vs. private groups where members know each other), our study reveals noteworthy disparities in the nature of corrections and responses. These variations can be attributed to the prevailing group norms and the level of familiarity among members. They point towards the existence of underlying power dynamics and trust within the community.

While previous studies have made significant contributions to understanding social corrections ([Allen, Arechar, and others2020]), specifically on WhatsApp ([Ng and Neyazi2022, Malhotra and Pearce2022, Pasquetto et al.2022]), they also exhibit certain limitations that warrant further investigation. Many of these studies rely on surveys or qualitative/interview based approaches, limiting their ability to provide a comprehensive quantitative analysis of the phenomenon. By contrast, our paper offers a robust quantitative examination of community-driven fact-checking, by making use of two different large scale datasets which capture natural, in-the-wild fact-checking behavior, which allows for more reliable and generalizable findings.

Considering the widespread misinformation on WhatsApp ([Resende et al.2019]) and the challenges posed by content moderation due to encryption, community-driven efforts emerge as a potential solution to combat this problem. Notably, this work represents the first exploration of fact-checking in a non-US context within real-world settings, making the obtained insights and datasets valuable to the research community. Given WhatsApp's status as the most downloaded app on Android and the significant amount of time users spend consuming information on the platform, developing methods to timely identify and debunk rumors before they spread further can greatly enhance user experiences, considering the crucial role of timing in the fact-checking process (Brashier, 2021).

By analyzing our findings and datasets, further advancements can be made in the area of community-driven fact-checking, including identifying the most suitable fact-checkers, exploring methods to incentivize their efforts, and providing them with appropriate tools for efficient fact-checking. Additionally, this research will contribute insights on the contextual factors that influence the effectiveness of fact-checking, addressing the question of how to conduct effective fact-checking.

2 Related Work

Prior research has examined the effects of fact-checking, though less often by fellow community members (cf. [Friggeri et al.2014]), and typically found it effective ([Wood and Porter2019, Bode and Vraga2018, Vraga and others2018]). In this literature, we find that community-based fact-checking, encrypted platforms like WhatsApp, and countries in the global south are comparatively understudied. In particular, we think a focus on community fact-checking on WhatsApp in India is warranted for three main reasons: (i). WhatsApp is used by over 500 million Indians, making up a huge portion of WhatsApp/Facebook's market base, and yet is highly understudied. (ii). WhatsApp is usually considered personal communication, and hence information consumed on WhatsApp is highly trusted ([Banaji and Bhat2019]). This makes community driven fact-checking more plausible, and (iii). WhatsApp has little ability to moderate content due to encryption. Community-driven solutions, including automatically identifying optimal allocation of resources for fact-checking by users might be the way to go. Moreover, the dialogue nature of the conversation makes it an extremely natural way to fact-check content.

Recently, there have been efforts for community driven fact-checking on Twitter which is promising ([Allen, Arechar, and others2020, Resnick and others2021, Pröllochs2021, Appelman and Leerssen2022]). There is also a rich body of work in understanding community moderation on platforms like Reddit ([Jhaver and others2019]). Our study is different from these works, as it provides insights on fact-checking on a class of platforms that are understudied in the literature on social media and politics: messaging applications. These applications, particularly WhatsApp, are especially popular in developing countries. WhatsApp has several fundamental features (such as end-to-end encryption, popularity in the global south, mobile-first app, etc) that distinguish it from well-studied social media platforms like Facebook, Twitter or Reddit. Community moderation on platforms like Twitter (community notes) or Reddit is very different from our setup since on those platforms, you can get corrected by strangers (or members who are losely associated with your interests). This dynamic is very different than a community like WhatsApp, which mostly consists of people you might know personally in real life. This leads to a completely different dynamic, as our results point out. Moreover, most moderation on large social media platforms like Facebook or Twitter is usually top-down, where platforms enforce rules and make decisions on behalf of the posters. This

paper explores a bottom-up approach, since a top-down approach is not feasible on a platform like WhatsApp due to encryption. As the world moves towards encryption and privacy,[1] understanding and developing tools that enable bottom-up solutions which respect the privacy of the users are the need of the hour.

Several studies have examined the phenomenon of social correction on WhatsApp in different contexts. [Kligler-Vilenchik2022] study a WhatsApp group in Israel managed by a journalist and study the concept of 'collective social correction' in the group. [Varanasi, Pal, and Vashistha2022] explored the role of gatekeepers in correcting misinformation among rural and urban WhatsApp users in India, highlighting the prevalence of corrections through discussions with trusted sources. They also discussed the nuances of soft nudges versus direct confrontation and the influence of authority in correction within polarized settings. [Ng and Neyazi2022, Malhotra and Pearce2022, Vijaykumar et al.2022] used mixed methods studies in Singapore and Brazil to identify three social correction strategies on WhatsApp: correction to the group, correction to the sender only, and no correction. [Rossini et al.2021, Rossini2023] compared social corrections on WhatsApp and Facebook, finding that WhatsApp users were more likely to engage in, experience, and witness social corrections compared to Facebook. Finally, [Pasquetto et al.2022] investigated the social debunking of misinformation on WhatsApp, emphasizing the effectiveness of corrections from in-group members. This is particularly relevant to our case as all members belong to the same political party.

[Bode and Vraga2021] focuses on the effectiveness of "observational correction," where people adjust their beliefs after witnessing corrections made to someone else's misinformation on social media. The paper suggests this method is scalable and populist, relying on ordinary users to correct misinformation. Our research adds to this burgeoning field by focusing on community-driven fact-checking within the encrypted environment of WhatsApp. Unlike [Bode and Vraga2021], which discusses the populist nature of corrections on public platforms, our research addresses the unique challenges posed by WhatsApp's encrypted setup. We reveal that even in a more private, encrypted environment, community-driven corrections are infrequent, and when they occur, they are often ignored or met with resistance. [Malhotra and Pearce2022] Investigates the role of politeness and relational norms in misinformation correction within familial WhatsApp groups in India. The paper shows that people employ indirect strategies to correct misinformation while adhering to cultural norms of respect and politeness. [Pearce and Malhotra2022] Uses affordances perspective and face-negotiation theory to study how different social and mobile media affordances influence channel selection for misinformation correction within family groups. It emphasizes that corrections rarely occur on group chats, attributing this to a complex interplay of social and contextual factors. [Malhotra, Scharp, and Thomas2022] Analyzes online posts related to misinformation correction and looks at how the meaning of 'misinformation' and 'correctors' varies depending on social contexts and norms. The paper extends Relational Dialectics Theory by exploring the complexities of interpersonal misinformation correction. Our research complements this by examining the dynamics of power and trust within WhatsApp groups, shedding light on why corrections are or aren't effective. While these two works delve into relational norms and the role of politeness in misinformation correction, they focus on specific cultural contexts and familial groups. Our study adopts a broader view, investigating a range of public and private WhatsApp groups with varying levels of intimacy among members. Unlike these papers, we provide quantitative insights, enabling us to make generalized observations about the frequency and types of corrections on WhatsApp.

Overall, our research not only provides a novel quantitative lens to explore community-driven fact-checking on WhatsApp but also unveils critical dimensions previously unexamined in the literature, such as the nuanced role of administrators and the tone of fact-checking. Further, we probe the underlying dynamics of power and trust within groups, offering a refined understanding of their impact on the efficacy of fact-checking. Our findings have the potential to inform the design of more effective, user-centric reporting interfaces and can be leveraged to craft policies and technology solutions aimed at mitigating misinformation on encrypted platforms.

3 Datasets

We use two datasets: (i) Public WhatsApp groups: The dataset was collected by [Garimella and Eckles2020], which included over 5,000 public groups scraped from the web and social media. These groups were either created by the political parties themselves or by supporters and posted large quantities of misinformation. Public WhatsApp groups discussing politics are quite popular and widely used by political parties in India and Brazil to reach potential supporters ([Lokniti2018, Newman et al.2019]). The data was collected during the national election in India in 2019. (ii) Private political WhatsApp groups: These groups were collected by [Chauchard and Garimella2022] and include data from over 400 private political groups. The groups were identified during a survey on-ground where the authors requested admins of the groups for access to political party managed groups. The dataset was collected over a 9-month period spanning late 2019 and early 2020 and took place in the state of Uttar Pradesh in India. Both the datasets were collected by previous works, which had gone through institutional review at their institutions. No easily identifiable information (phone numbers) were present in the data we analyzed, thus making it difficult to link the users back to their messages. That said, datasets containing conversations are never easily anonymizable. We tried our best to make sure all the precautions were taken while handling the data.

From this point forward, we will refer to the two datasets

[1] e.g. https://tcrn.ch/3OeOb94/ Facebook is bringing end-to-end encryption to Messenger calls and Instagram DMs

as the Public dataset and the Private dataset, respectively.

The two datasets provide a comprehensive representation of a diverse environment, encompassing various aspects such as user composition, user acquaintance, and message discipline. The Public dataset primarily consists of groups comprising strangers who have joined or been added to a group due to their support for a political party. These groups are publicly accessible, allowing anyone to join and contribute messages, resulting in dynamic and volatile discussions on a wide range of topics ([Freelon2017, Garimella and Eckles2020]). On the other hand, the Private dataset comprises groups predominantly composed of party workers who actively share and forward party-related content on WhatsApp. The group members are typically acquainted with each other, fostering a sense of familiarity within the group. Consequently, the messaging within these groups revolves primarily around local politics and related discussions ([Chauchard and Garimella2022]).

These two datasets offer excellent opportunities for studying community-driven fact-checking due to their unique characteristics. In the Public dataset, the diverse user composition consisting of strangers with varying backgrounds and perspectives creates a rich environment for fact-checking discussions. With users joining groups based on their political affiliations, the dataset captures a wide range of viewpoints and ideologies. With an estimated 100 million WhatsApp users being a part of at least one such public group discussing politics in India ([Lokniti2018]), it is interesting to study whether this diversity of viewpoints and the lack of a need for maintaining any reputation fosters robust debates and critical analysis, enabling fact-checking efforts to benefit from the collective intelligence and knowledge of the community. In contrast, the Private dataset's composition of predominantly acquainted party workers provides an intriguing setting for studying fact-checking dynamics within a more closed and cohesive community. The familiarity among group members facilitates effective collaboration, trust-building, and information sharing. Within this context, fact-checking efforts might be more streamlined and targeted, benefiting from the strong social ties and a sense of responsibility and accountability among the participants.

Identifying corrections. To identify corrections to misinformation, we used the keyword 'fake' (and slight variations like 'fack','feke','feck') and translations of it in various Indian languages: ಅಬದ್ಧ , తప్పు , ఫేక్ , फ़ेक, झूट, अफ़वाह, ফেক , ફેક , ಫೇಕ್ , ଫେକ , ਫੇਕ , பீக் , فیک on our datasets. The choice of the keyword "fake" was based on an initial exploratory analysis of the datasets, where we observed a high frequency of its use in corrections. We started by looking for responses to hundreds of images annotated as misinformation (for a different paper (citation anonymized), on the WhatsApp datasets). These hundreds of misinformation images were identified by annotating a random sampling of all content shared in our groups over a period of one month. Since we started with a random sample, and qualitatively looked at *all* the responses to misinformation content, we were relatively sure that there were no other significant variants used to report or correct misinformation. Having included the different local language versions covers almost 99% of the cases of corrections we find manually. The cases we miss were a handful where a user replied sarcastically to misinformation with emojis or other memes.

For the Public dataset, there were over 1,986 results for these search keywords, and for the Private dataset, there were 123 results. In many cases, users on the WhatsApp threads posted that a message was 'fake' if they did not agree with the contents of the message. To filter out such cases, human annotators who were well versed with Hindi, Telugu and English went through all the search results for these keywords and manually classified whether the messages contained a real correction or not. They used Google translate for the small number of cases where the messages were in Urdu, Gujarati or Kannada. The annotators were also well versed with the local politics in India and had tools to fact-check content. They were provided access to an interface where the anonymized chat data was shown so they could get the entire context in which a message was posted. They started from a message which contained the 'fake' keywords (the 'correction' message) and checked which message it was referring to (the 'original' message). Next, they verified the veracity of the original message. If the original was really misinformation, they labeled the correction message and the original message and the users who posted these as 'poster' and 'corrector' respectively. In case the annotators could not identify the veractiy of the original message,[2] they ignored the message.

After filtering the false positives, we were left with 1,250 and 84 cases in the Public and Private datasets respectively where users explicitly reported something to be fake. Table 1 shows a summary of the two datasets.

Privacy. Given the private nature of WhatsApp, privacy was of paramount importance to us. The original studies where the datasets were obtained from had undergone institutional review prior to the data collection. Our study which makes use of subsets of the datasets also went through institutional review from (anonymized) university. The datasets were anonymized and all identities of participants like names, emails and phone numbers were replaced with random identifiers using the Google data loss prevention library ([Cloud2023]) before working on the annotation.

4 Analysis

In this section, we start our analysis with studying how prevalent community-driven fact-checking is. Next, we focus on who fact-checks whom, what is fact checked and how, and finally whether community-driven fact-checking is effective.

[2]e.g. in case it was not fact checked anywhere and the annotators were not familiar with the topic or if they could not deduce the context in which the message was posted. This happened only for less than half a dozen instances in both datasets.

Table 1: Characteristics of our datasets

	Public	Private
#Groups	725	417
Public Correction	✓	✓
Strangers	✓	×
#Total Users	87,013	30,473
#Languages	6	2
#Misinfo annotations	partial	complete
#Posts corrected	1,250	84
#Posters	862 (1.09%)	330 (1.09%)
#Correctors	860 (1.08%)	78 (0.2%)

4.1 How often is this done?

Social correction messages comprise approximately 0.06% of all messages in the Public dataset when considering them as a fraction of the total messages. However, this measure alone does not provide an accurate assessment of their prevalence. A more meaningful comparison can be made by considering the correction messages in relation to misinformation posts. In the Private dataset, the authors annotated a random sample of 10% of the dataset's images to identify misinformation. This allows us to estimate that approximately 11% of the misinformation posted in the Private dataset was corrected by community members. This finding suggests that while social corrections are infrequent overall, they do occur in a small subset of cases. The occurrence of social corrections on WhatsApp is influenced by the chat context, as indicated by the qualitative study conducted by [Ng and Neyazi2022, Malhotra and Pearce2022].

In terms of the number of users engaged in the process of social correction, roughly 1% of the users engage in posting misinformation and in correcting it in the Public data. However, in the Private data, most of the corrections are done by a much smaller number of users. This is expected, as in a private group, there is much more message discipline and admins enforce more control. We can see from Figure 2 that almost 80% of the users post or correct only one piece of misinformation. There are a handful of users who do this more often, some even doing it over 30 times.

Interestingly, there is zero intersection between correctors and posters of misinformation in both datasets. This suggests that the act of fact-checking and correction requires a certain level of motivation, knowledge, or inclination that may not be present in all group members. It could also indicate a lack of awareness or concern about the spread of misinformation among those who are actively sharing content. Understanding this division between correctors and posters has implications for designing effective strategies to combat misinformation on WhatsApp. It highlights the importance of identifying and engaging individuals who are willing and capable of fact-checking and providing them with the necessary resources, tools, and incentives to carry out this crucial task. Additionally, it suggests the need for targeted interventions to raise awareness and promote responsible information sharing among those who primarily act as posters, potentially encouraging them to participate in the correction process as well.

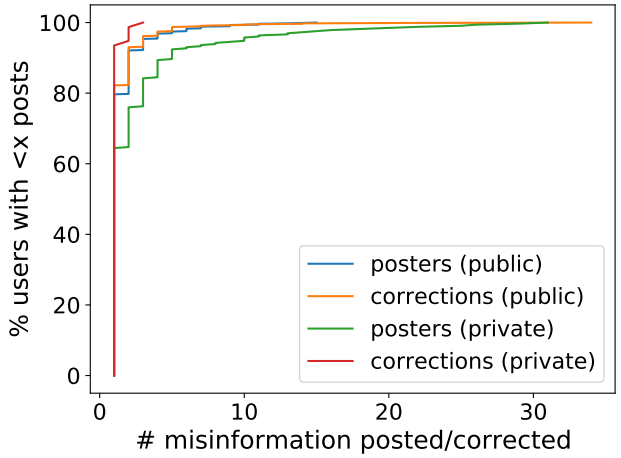


Figure 2: CDF of posting/correction. We see that over 80% of the users post/correct only once.

4.2 Who fact checks whom?

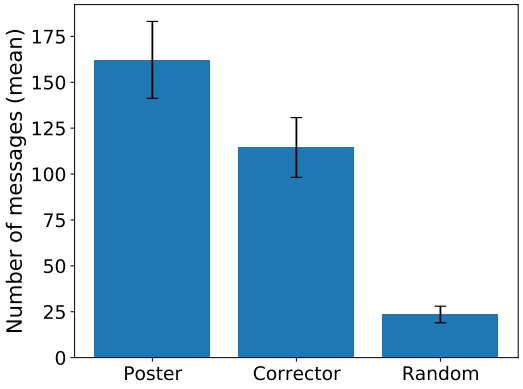
In this section, we try to understand who posts misinformation and who corrects them.

Activity. We start by measuring the activity levels of the posters and correctors and compare it with a random sample of users (of the same size as posters) in our dataset. Figure 3 shows the activity of these sets of users. We see that both posters and correctors are significantly more active than a random user in our datasets. It is surprising that the level of activity is significantly higher compared to random users since this indicates that these users are active members of the community. Notably, the activity level of posters tends to be slightly higher than that of correctors. This observation aligns with the expectation that the inclination to post misinformation may also be influenced by higher activity levels.

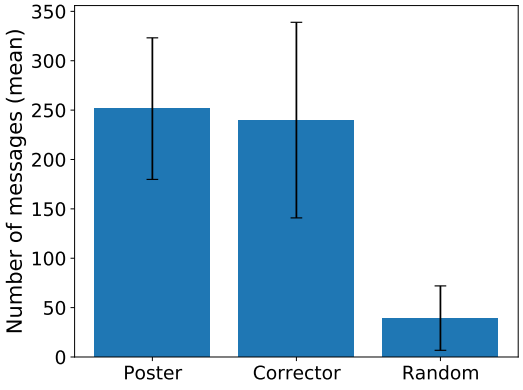
To better understand the relationship between user activity and the posting or correction of misinformation within WhatsApp groups, we conducted an analysis by categorizing users based on the total number of messages they posted in each group. The users were divided into five bins, representing each 20^{th} percentile of activity. We then calculated the percentage of misinformation posted and corrected by users in each activity bin.

The results, depicted in Figure 4, provide compelling evidence that the individuals who actively participate in the groups by posting messages are also the ones who play a significant role in spreading and rectifying misinformation. This finding challenges the assumption that those responsible for misinformation dissemination may be outsiders or peripheral members who contribute minimally to the conversation. Instead, it suggests that the most active users within the group are the primary contributors to both the propagation and correction of false information.

Moreover, a notable discrepancy emerges when comparing the private and public datasets. In the Private dataset, users in the most active bins account for approximately 10% of the messages that contain misinformation, while in

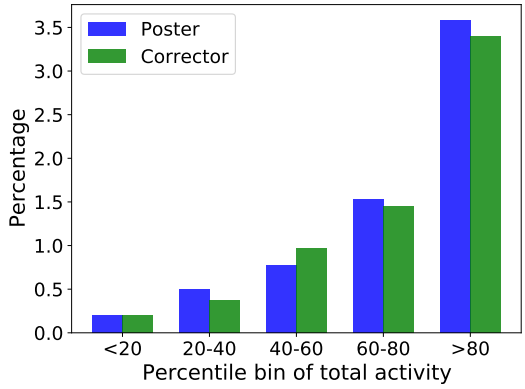


(a) `Public` dataset

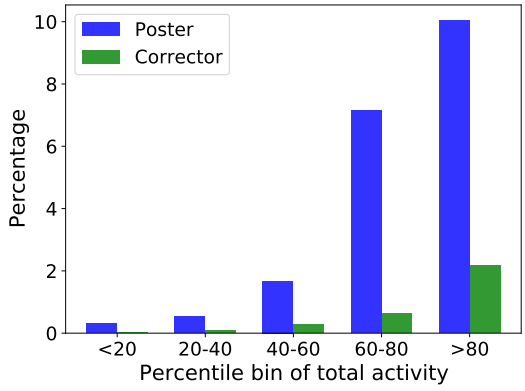


(b) `Private` dataset

Figure 3: Total posts by posters, correctors and random users. Error bars indicate 95% confidence intervals.



(a) `Public` dataset.



(b) `Private` data

Figure 4: Activity in a group vs. correction percentage.

the `Public` dataset, this percentage remains below 4%. In light of these findings, it becomes increasingly crucial to recognize the influence of active participants in both the propagation and correction of misinformation within WhatsApp groups. Strategies aimed at combating misinformation should focus on engaging these active users, as they hold significant sway over the information flow. By leveraging their involvement, interventions can effectively target the dissemination of false information and empower these key participants to play a vital role in mitigating its spread. Furthermore, understanding the nuances between private and public groups sheds light on the contextual factors that contribute to the dynamics of misinformation propagation, enabling more tailored and effective interventions to combat the issue.

Role of admins. We also examined whether the individuals who posted and corrected misinformation within their respective groups were administrators (admins) of those groups. Figure 5 illustrates the stark contrast in values between the `Public` and `Private` datasets. In the private dataset, admins are responsible for approximately one in every five corrections, indicating their active involvement in rectifying misinformation. Despite making a similar number of posts, admins demonstrate a greater commitment to correcting misinformation within private groups. It is noteworthy that a significant number of admins in both datasets are also found to post misinformation. Intriguingly, our data also reveals a more complex picture: administrators in both public and private settings are not only active in correcting misinformation but are also contributors of misinformation themselves. This complicates the general perception of admins as solely gatekeepers of truth and challenges us to rethink the dynamics of authority and credibility within these digital communities.

These findings imply that the responsibility for fact-checking does not rest solely on admins but extends to other members of the group as well. While admins play a role in both posting and correcting misinformation, the collective effort of the community is instrumental in combating false information within WhatsApp groups. This highlights the collaborative nature of misinformation correction, with multiple individuals actively participating in the process. Therefore, leveraging the community's collective knowledge and engagement becomes paramount in effectively countering misinformation.

Strength of the relationship between posters and correctors. The analysis until now looks at posters and correctors as independent users. In the next line of analysis, we test whether the relationship between the poster and corrector plays a role in the social correction. To quantify this, we

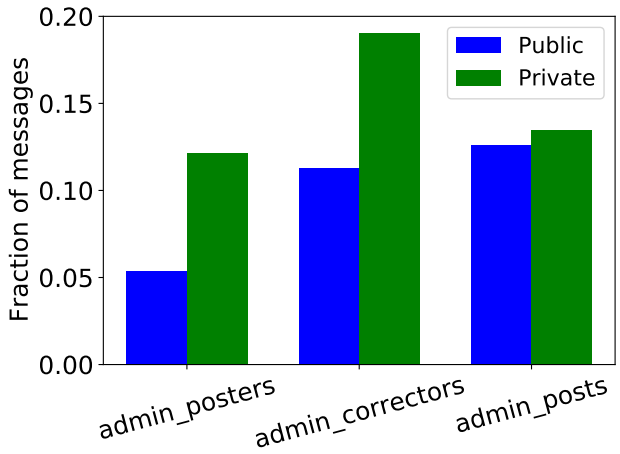


Figure 5: Fraction of messages posted by admins

created the notion of tie strength between users and measure whether this tie strength affects propensity to correct misinformation. We computed the tie strength based on how many times one user replies to another. We created an interaction graph, which is a directed graph of the users where an edge exists between two users if they posted a message one after the other and the weight of the edge indicates the number of interactions. Figure 6 plots this weight (or tie strength) between a poster-corrector pair and a random pair of users. The figure shows that most poster-corrector pairs have significantly stronger ties compared to a random set of users.

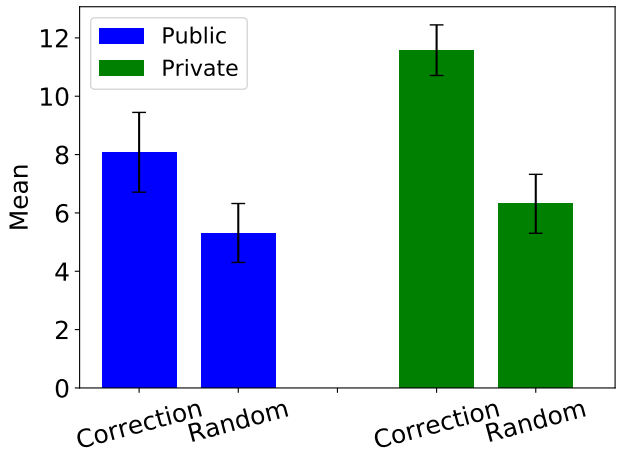


Figure 6: Tie strength between correctors and random users. 'Correction' is the tie strength between poster and corrector. Error bars indicate 95% confidence intervals.

Next, we measure the centrality of users in this interaction graph. We computed the Pagerank of the users in this graph. Intuitively, the Pagerank in the interaction graph indicates how important the user has been in terms of being replied to. We find that in line with Figure 3, posters and correctors are central members of the community. Interestingly, for the `Public` dataset, posters have a significantly higher Pagerank compared to correctors, which is not the case for the `Private` dataset.

This finding indicates that the relationship between posting behavior, influence, and correction activity can vary depending on the social context and the level of openness of the group. In public groups, where information is shared with a larger audience and there may be a higher level of competition for attention, individuals who actively post messages, even including misinformation, may gain more prominence within the network. On the other hand, in private groups, where trust and familiarity are more pronounced, the influence and importance of individuals may be determined by factors beyond their posting behavior alone.

Overall, Figures 5 and 7 compare both explicit (being an admin) and implicit (high centrality) notions of power and show that users who post misinformation have both implicit and explicit power in the groups.

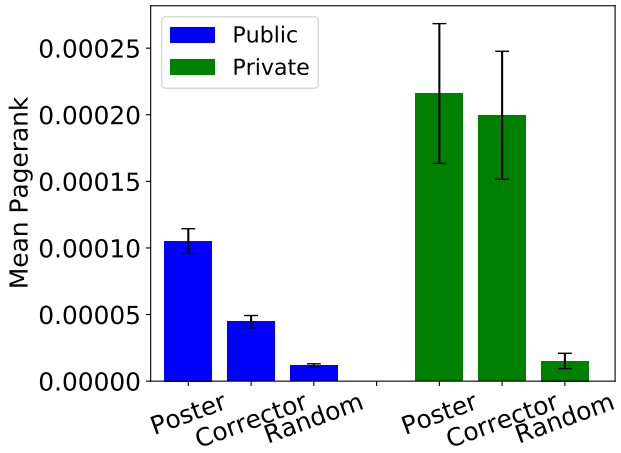


Figure 7: Pagerank of posters, correctors and random users.

Predicting posters & correctors. In this section, we build a classifier to predict given a user whether they are a poster, corrector or a random user. We used 4 simple features: the centrality of the user in the interaction network, whether they are an admin, number of messages posted, the average response time before their previous message (captures how active and attentive the use is in the group, represented by t_δ). We use only these metadata features (and not the content of their messages) so that it is realistic in an end to end encrypted setup. We set this up as a three class prediction task predicting whether a user is a poster, corrector or a random user. A random guess would give an accuracy of 33%. For the three class prediction, a Random Forest classifier performs much better than random with an accuracy of 47.8%. The results are summarized in Table 2 for various classification tasks. The idea behind this exercise is not to build a perfect classifier to predict a user type but to show that these seemingly simple metadata signals contain value in making predictions of the type of users.

4.3 What is fact-checked and how?

In this section, we qualitatively analyze the content that is fact-checked and the types of social correction. Given the

Table 2: Accuracy for the prediction of user types

	Accuracy	Random Accuracy
Three Class	0.47	0.33
Corrector vs. Random	0.646	0.5
Poster vs. Random	0.719	0.5
Poster vs. Corrector	0.579	0.5

multi modal nature of the data, quantitative analysis is difficult. Hence, we chose a mixed-methods approach where we qualitatively look through the data and compute statistics of interesting findings.

Type of misinformation. Depending on the dataset, the type of misinformation that was corrected was different. In the `Public` dataset, most of the misinformation was rumors regarding the elections, issues targeting their own party (misinformation targeting the other parties often goes unquestioned), and spam/scams/forward bait. Qualitative comparison of the misinformation content that is fact checked vs. not indicates that there is preference to non-dissonant content.

Because we had annotations for misinformation in the `Private` dataset, we can study what happens when misinformation is posted and what types of content is picked to be corrected. As we saw in Section 4.1, only 11% of the misinformation posted in the private data was corrected. Only a small fraction of the corrections involve the posting users being warned about consequences for posting misinformation. The remaining 89% of the cases where the misinformation is not corrected are mostly because the misinformation was ideologically congruent to the group (we could only find 3 instances of ideologically congruent misinformation being corrected), or because it got drowned out by other content which was posted later. We often find that (mis)information is posted enmasse with a single user sometimes posting half a dozen pictures/videos. In such cases if someone does not correct the misinformation immediately, they are not corrected. The corrections mostly increase at times of high activity periods, when there is a chance of misinformation. These peaks correspond to events like the terror attack in Kashmir and the retaliation strikes by India, the 2019 election and the results being announced.

Types of corrections. We manually went through all the corrections in both datasets and coded them into different categories. The categories were arrived at through an iterative process where the authors first made a pass over the data to collect rough examples of categories which were then refined to come up with a final set of categories. Note that a single correction can belong to multiple of these categories. Figure 8 shows the distribution of the four most prominent categories.

Some interesting observations emerge: In both the datasets, the most prevalent type of correction is politely correcting the user who posted misinformation. Some examples of polite corrections:[3]

> 'This is fake. Plz dont send these type of msg' (sic)

> 'This is fake brother. Please know if it is true before posting'

> 'Dont post fake news like this brother please.... '

> 'I really wish it was true but it is fake brother'

> 'Boss this is all fake. Please thing your self' (sic)

> 'Sir this is fake'

The second most popular category of corrections was just users saying one or two words calling a post as 'Fake' without providing any context. Given the vast amount of messages these groups receive, such single word corrections almost always get lost in the deluge of messages.

The third category we coded were corrections where the correctors insult the poster for posting misinformation.

> 'You are a liar and your posts are full of lies'

> 'You are an idiot for believing these scams. You always post fake news'

> 'Can't you just read what is written above the picture? FAKE NEWS 😡'

> 'It's fake. Do you think we are idiots?'

> 'Guys this user is making us look like fools. Please do not forward this to your groups'

Finally, the fourth category we found was corrections which provided evidence along with the correction. Some users (correctly) claimed that the images being shared were edited since they contained (badly) photoshopped images. Interestingly, evidence pointing to popular fact-checking websites was extremely rare. We could find less than half a dozen instances of links to fact-checking websites.[4]

> 'Brother this is a fake photo. The BJP IT cell is trying to ruin the reputation of our leader Rahul Gandhi. You can see that the photo is edited. Check above the Rahuls head.'

> 'Fack news this not Ambur.. This near Dindigul' (sic)

> 'Fake circulating for last 3 years.. <link to a fact checking site>'

> 'Fake and photoshop... <image showing the photoshop location>'

[3] In the examples shown, the text has been translated to English and the names (if any) were changed to avoid identifying the users. Some of the text was left as is in case it was mostly in English.

[4] This is surprising and interesting because according to the Duke Reporters Lab report on fact-checkers, India has the most number of fact-checking agencies in the world with over 25 registered fact-checking agencies ([Lab2023]).

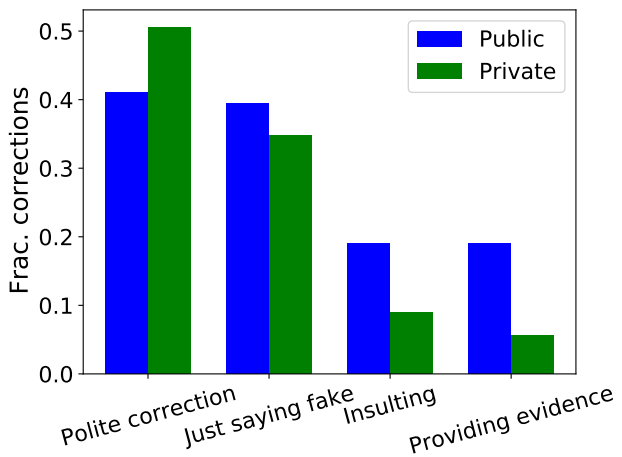


Figure 8: Distribution of correction types in `Public` and `Private` datasets.

> 'Fake. Rahul meeting in UP, I guess. Not a Modi meeting'

A comparison between the correction types across the two datasets reveals interesting observations. In the `Private` dataset, the fraction of polite corrections are significantly higher, with almost half of the corrections being polite. On that note, the fraction of insulting corrections is also highly rare. This could be because of the private and intimate nature of the group where the participants know each other. Surprisingly, the fraction of corrections which provide evidence for the corrections are also very rare in the `Private` dataset. This could be because of the intricate power relationships which exist in the groups and users avoiding being too confronting ([Ng and Neyazi2022]).

Other interesting cases emerged during the annotation, which show the intricacies involved in the group dynamics. Sometimes even though the message was fake and was already fact checked (externally), users got upset that they were corrected. This exchange shows a back and forth where the user seems to be hurt and wants get an explanation on why he was corrected. Even though they might believe that the information is false, they are questioning the premise of the correction.

> User 1 (Corrector): 'I think is fake'
>
> User 2 (Poster): 'Ansari ji, why are you saying that? what do you achieve?'
>
> User 2: 'answer me'
>
> User 2: 'why do you say its fake? I saw it on TV'
>
> User 1 (Corrector): 'I did not see anything on TV'
>
> User 3: 'I thought it was fake too'

There are instances where users quit the groups because of a lack of proper moderation. For instance:

> User1 (Poster): 'BREAKING NEWS: Sad news *Former Finance Minister and senior BJP leader Arun Jaitley passed away*'
>
> User 2 (Corrector): 'Everyone, this is fake news'
>
> User 3: 'I do not see anything on the TV news.'
>
> User 4: 'I am leaving this group bcoz there are too many users who post fake news.' (sic)
>
> User 4: left chat
>
> User 3: 'You are right'
>
> User 2 (Corrector): 'I am leaving tooo'

In some groups where admins do not take action, users explicitly reply to every message posted by a scam/spam account labeling them as fake and asking the admin to take action. Sometimes people request admins to take action. Since admins do not care in many groups. In some cases, admins do act and remove members who post misinformation (or even true information which is ideologically incongruent).

> 'Admins, please remove these guys who post fake news ...'

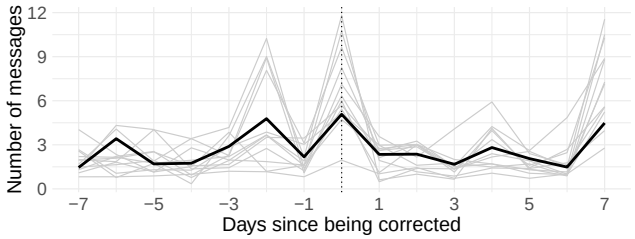
We also see messages of people warning their fellow group members not to forward the content the correctors flagged as fake:

> 'Do not spread the fack news' (sic)
>
> 'Don't open its fake'
>
> 'This one fake don't forward other group' (sic)

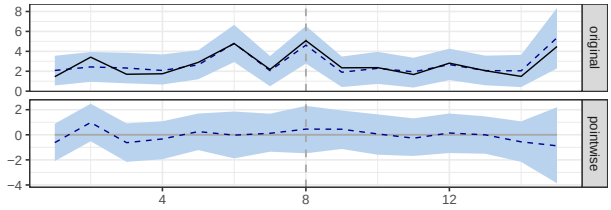
We observed that the type of reaction to a fake post mostly depended on the type of poster: if it is a 'bad' actor (either users considered outsiders, users who post misinformation deliberately or spammers), the correction is mostly either a one word 'Fake news' or insulting. If the poster is an active member of the group, there is usually a personal and polite response and in some cases evidence is provided and a follow up conversation ensues.

4.4 Is community fact-checking effective?

We finally try to answer whether correcting a user changes their behavior in any way, including their posting patterns. We treat it as a regression discontinuity with the time the correction happened as the 'treatment' day. We measure the impact of the correction of three variables: the number of posts made by the user before/after the correction, the rate of engagement with the community before/after and the amount of misinformation posted before and after. Let's first consider the case of the number of posts. We look at the number of posts made by the user one week prior to the correction and one week after the correction. We use the 'CausalImpact' library in R ([Brodersen et al.2015]) to identify the impact. The library builds synthetic controls ([Abadie2021]) based on a weighted average of 'untreated' subjects. These untreated subjects are a random sample of users for whom a random 'treatment' time is selected and the messages counted one week before and after.



(a) The dark line indicates posting activity one week before and after being fact checked. Grey lines indicate random samples used for synthetic control.



(b) Top: Volume and synthetic control fit, Bottom: point wise difference. We can see that there is no significant difference to the control set.

Figure 9: The effect of social correction on volume of posts over time (`Public` dataset).

Figure 9 shows the timeseries of posts one week before and one week after being corrected. The dark line in Figure 9a plots the number of messages posted per day. The grey lines indicate ten sets of random samples. The top plot in Figure 9b shows the volume and the weighted average from the synthetic control estimate. The pointwise difference (Figure 9b) shows the pointwise difference between the control and treatment. If there was a causal effect, the pointwise difference should be significantly different than zero post treatment. We see that there is no significant change, indicating that there is no effect on the number of messages posted by the user after being corrected. We performed a similar analysis looking at any changes in the engagement they have with the community by measuring the t_δ: time taken to respond to a message in a group before and after being corrected. We also find no evidence of any change in the t_δ. These effects are consistent for both the `Public` and `Private` datasets. We omit plots for time delta and for the `Private` dataset.

On the `Private` dataset, we also measured whether the users who got corrected posted any less misinformation going forward. We do not find any change in their posting behavior, with 62% of the users posting at least one misinformation after being corrected.

5 Discussion

This paper offers a comprehensive analysis of community-driven fact-checking based on two extensive real-world datasets. In contrast to prior research in this domain, our quantitative approach enabled us to address key questions pertaining to the frequency, individuals involved, their attributes, and the overall effectiveness of community-driven fact-checking efforts.

The findings highlight several important insights regarding community-driven fact-checking on WhatsApp. Firstly, we find that only less than 1% of the users in a group engage in the correction and most users only do it once. The low prevalence of community-driven corrections suggests that the practice is not yet widespread within these platforms, especially in a *public* setting. This could be attributed to various factors, such as limited awareness about the importance of correcting misinformation, lack of incentives or mechanisms to encourage corrections, or even social dynamics that discourage challenging misinformation in close-knit groups.

The study also reveals that the responsibility for fact-checking falls primarily on active community members, rather than group administrators. This highlights the potential role of influential users within the community in shaping the discourse and promoting accuracy. However, it is worth noting that relying solely on community members for fact-checking may pose challenges in terms of scalability and ensuring the accuracy and reliability of the corrections. Future research could explore strategies to incentivize and empower community members to take on a more active role in fact-checking, while addressing potential biases.

The observation that users often do not engage in social-correction and may even disregard or feel offended by public corrections raises questions about the effectiveness of community-driven fact-checking efforts. It appears that the mere presence of corrections may not be sufficient to prompt individuals to revise their beliefs or correct their misinformation. This finding aligns with prior research on cognitive biases and the challenges of changing deeply held beliefs, suggesting that alternative approaches beyond direct correction may be needed to effectively combat misinformation in these contexts.

Furthermore, the significant variations in correction types and responses across different groups underscore the importance of considering group norms and dynamics when designing interventions for fact-checking. The observed differences between public and private groups suggest that social factors, power dynamics, and group cohesion play a crucial role in shaping the acceptance and effectiveness of corrections. Understanding these contextual nuances can inform the development of targeted interventions that align with the specific dynamics of each group, facilitating more effective fact-checking practices.

Practical implications. Our work is not just academic; it has actionable implications. Our findings can directly inform the design of more effective fact-checking mechanisms, reporting interfaces, and overall user experiences in both encrypted and public platforms.

(i) Designing Adaptive Fact-Checking Tools: The observed variations in community responses to corrections—rooted in group norms and familiarity among members—suggest that one-size-fits-all fact-checking tools may not be effective. Tailored tools that adapt to specific community norms could be developed. For instance, platforms can enable users to earn 'trust badges' based on their history of accurate fact-checking, amplifying their corrections within the group. Platforms could also develop user-friendly tools and interfaces within encrypted chat platforms to facilitate

community-driven fact-checking, such as providing automated fact-checking suggestions to users when they encounter previously debunked false information, empowering them to make informed judgments.
(ii) Enhanced Reporting Interfaces: Considering the limited role of administrators in corrections, platforms can also introduce a tiered moderation system where trusted members share fact-checking responsibilities. This decentralizes the power structure, potentially making corrections more palatable. These designs could establish clear reporting mechanisms that allow users to submit misinformation for fact-checking quickly in a way they that may not compromise their reputation in their group/community.
(iii) User Experience and Notification Design: We discover that individuals are often apathetic or even offended by corrections suggests that how the message is delivered matters. By testing on different types of correction notifications–varying in tone and format–platforms can determine which are most effective in prompting action without causing offense. They could also consider the influence of group norms and tailor interventions to align with the specific norms of each community.
(iv) Incentive Structures for Active Participation: Our paper reveals that a bulk of the responsibility for fact-checking falls on a few active members. Implementing incentive mechanisms to encourage broader community participation in fact-checking could be beneficial. This could be done by, for instance, introducing a reputation system could motivate more users to contribute to a more accurate information ecosystem; or by providing incentives for users who actively engage in (accurate) fact-checking, such as badges or recognition within their communities.

Limitations. While this study provides valuable insights into community-driven fact-checking in encrypted chat platforms, it is not without limitations. The reliance on observational data poses challenges in establishing causal relationships between corrections and behavior changes. Observational behavioral data, though powerful, can not be used to answer certain types of social science questions involving psychological traits of the users involved such as: what would the users who are fact-checked publicly in the group feel and react? Who are the fact-checkers? What encourages users to actively fact-check? The observational nature of the data also limits certain angles of qualitative study, such as why users may choose to (or not) respond to a message. 22% of the groups in our `Public` data did not have any corrections. Over 55% of the groups had just one correction. There were some groups which had over 40 corrections. Understanding these difference between the groups can not be done using our setup. Future research could employ experimental designs, interviews or survey methodology to enhance the robustness and generalizability of the findings.

A related issue is the inherent sample bias due to its observational methodology. The groups we investigated were either publicly accessible or had administrators who permitted our access. These conditions inherently introduce selection bias both in the types of users and the content that are represented. While it might be tempting to generalize our findings, caution is needed due to this sample bias. However, it is important to note that the prevalence of such public groups [Lokniti2018] and politically-affiliated groups [Perrigo2019] in India, as cited in previous studies, suggests that our dataset could still provide valuable, albeit non-trivial, insights into community dynamics. Moreover, the challenges associated with obtaining large-scale datasets from WhatsApp cannot be overstated, owing to encryption and stringent privacy measures. Thus, even a biased glimpse into the platform through millions of messages holds value, as it affords us an otherwise inaccessible window into community interactions on encrypted platforms. Given the critical nature of misinformation and its corrections, policy discussions have even ventured into the territory of dismantling encryption as a potential solution.[5] Research like ours contributes empirical evidence to inform these policy considerations, thereby grounding them in data-driven insights.

Another limitation of our set up is that we focus solely on public corrections, neglecting any private corrections that may occur (peer to peer, outside of our data collection). As noted by ([Ng and Neyazi2022]), it is important to acknowledge the potential existence of private corrections, particularly within the `Private` dataset (where the users know each other), which could potentially be more effective than public corrections.

Despite these limitations, our study serves as a robust foundation for discussing community-driven approaches to combat misinformation within the context of end-to-end encryption. Notably, marginalized communities in the United States often lack access to fact-checking resources, as they may not be active on mainstream platforms such as Twitter or Facebook ([Graves2018]). Given the persistent targeting of these communities with misinformation and disinformation ([Woolley2022]), adopting a community-driven model for correcting misinformation, informed by our research findings, can be particularly beneficial for them. Moreover, our findings have broader implications, guiding researchers, policymakers, and platform administrators in the development of targeted and effective strategies and interventions for fact-checking.

Broader impact statement. An important consequence of the convenience nature of our datasets is the bias in the datasets. For both datasets, the original authors ([Garimella and Eckles2020, Chauchard and Garimella2022]) also collect the demographics of the users in these groups. In both the datasets, the majority of the users (97%) were men. The age distribution was fairly uniform across different age groups, with the 25-34 age group being the most prevalent, accounting for approximately 30% of users in both datasets. Though all age groups were represented, our dataset is highly skewed towards men, which ignores how women and other genders would behave in such scenarios.

One key consideration in the idea of community-driven fact-checking is whether it is fair to rely solely on the community to combat misinformation. Community members

[5]https://www.theguardian.com/technology/2023/may/08/whatsapp-could-disappear-uk-over-privacy-concerns-ministers-told

may have varying levels of expertise, access to resources, and time availability to engage in fact-checking activities. This raises concerns about the burden placed on individuals within the community, particularly if fact-checking becomes a labor-intensive task that disproportionately falls on certain individuals or groups. It is crucial to consider the potential inequities and challenges that may arise when community-driven fact-checking is solely relied upon. For instance, a major concern is the potential for false or misleading corrections by a majority community, which can inadvertently perpetuate misinformation or harm particularly against minority groups. It is crucial to think through the accuracy, accountability, and transparency in the fact-checking process, with clear guidelines and standards for evaluating and correcting information.

Additionally, while community-driven fact-checking empowers individuals to take an active role in countering misinformation, particularly in an end-to-end encrypted content, it is important to acknowledge the role and responsibility of platforms. Platforms have a responsibility to provide reliable information and implement measures that mitigate the spread of misinformation ([Paris and Pasquetto2024]). Encryption and privacy might appear to be bottlenecks for the platform to be involved, but despite encryption WhatsApp does have significant control over how they handle content quality on the platform ([Jones2017]).

References

[Abadie2021] Abadie, A. 2021. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature* 59(2):391–425.

[Allen, Arechar, and others2020] Allen, J.; Arechar, A. A.; et al. 2020. Scaling up fact-checking using the wisdom of crowds.

[Appelman and Leerssen2022] Appelman, N., and Leerssen, P. 2022. On trusted flaggers. *Yale Law School: Information Society Project*.

[Arun2019] Arun, C. 2019. On whatsapp, rumours, lynchings, and the indian government. *Economic & Political Weekly* 54(6).

[Banaji and Bhat2019] Banaji, S., and Bhat, R. 2019. WhatsApp vigilantes: An exploration of citizen reception and construction of WhatsApp messages' triggering mob violence in india.

[Bode and Vraga2018] Bode, L., and Vraga, E. K. 2018. Correction of global health misinformation on social media.

[Bode and Vraga2021] Bode, L., and Vraga, E. K. 2021. People-powered correction: Fixing misinformation on social media. In *The Routledge Companion to Media Disinformation and Populism*. Routledge. 498–506.

[Brodersen et al.2015] Brodersen, K. H.; Gallusser, F.; Koehler, J.; Remy, N.; and Scott, S. L. 2015. Inferring causal impact using bayesian structural time-series models. *AAS* 247–274.

[Chauchard and Garimella2022] Chauchard, S., and Garimella, K. 2022. What circulates on partisan whatsapp in india? insights from an unusual dataset. *JQD: Digital Media* 2.

[Cloud2023] Cloud, G. 2023. Cloud data loss prevention.

[Freelon2017] Freelon, D. 2017. Campaigns in control: Analyzing controlled interactivity and message discipline on facebook. *Journal of Information Technology & Politics* 14(2).

[Friggeri et al.2014] Friggeri, A.; Adamic, L.; Eckles, D.; and Cheng, J. 2014. Rumor cascades. In *ICWSM*.

[Garimella and Eckles2020] Garimella, K., and Eckles, D. 2020. Images and misinformation in political groups: Evidence from whatsapp in india.

[Graves2018] Graves, D. 2018. Understanding the promise and limits of automated fact-checking.

[Jhaver and others2019] Jhaver, S., et al. 2019. User behavior after content removal explanations on reddit.

[Jones2017] Jones, M. 2017. How whatsapp reduced spam while launching end-to-end encryption.

[Kligler-Vilenchik2022] Kligler-Vilenchik, N. 2022. Collective social correction: addressing misinformation through group practices of information verification on whatsapp. *Digital Journalism*.

[Lab2023] Lab, D. R. 2023. Fact-checking across the world.

[Lokniti2018] Lokniti, C. 2018. How widespread is whatsapp's usage in india? Live Mint.

[Malhotra and Pearce2022] Malhotra, P., and Pearce, K. 2022. Facing falsehoods: Strategies for polite misinformation correction. *International Journal of Communication* 16:22.

[Malhotra, Scharp, and Thomas2022] Malhotra, P.; Scharp, K.; and Thomas, L. 2022. The meaning of misinformation and those who correct it: An extension of relational dialectics theory. *Journal of Social and Personal Relationships* 39(5):1256–1276.

[Newman et al.2019] Newman, N.; Fletcher, R.; Kalogeropoulos, A.; and Nielsen, R. K. 2019. Reuters Institute Digital News Report 2019 . Reuters Institute for the Study of Journalism.

[Ng and Neyazi2022] Ng, S. W. T., and Neyazi, T. A. 2022. Self-and social corrections on instant messaging platforms. *IJOC*.

[Paris and Pasquetto2024] Paris, B., and Pasquetto, I. 2024. Hidden virality and the everyday burden of correcting whatsapp mis-and disinformation. *Cambridge Studies on Governing Knowledge*.

[Pasquetto et al.2022] Pasquetto, I. V.; Jahani, E.; Atreja, S.; and Baum, M. 2022. Social debunking of misinformation on whatsapp: The case for strong and in-group ties. *CSCW* 6.

[Pearce and Malhotra2022] Pearce, K. E., and Malhotra, P. 2022. Inaccuracies and izzat: channel affordances for the consideration of face in misinformation correction. *Journal of Computer-Mediated Communication* 27(2).

[Perrigo2019] Perrigo, B. 2019. How Whatsapp Is Fueling Fake News Ahead of India's Elections. https://time.com/5512032/whatsapp-india-election-2019/. [Accessed 14-09-2023].

[Porter and Wood2021] Porter, E., and Wood, T. J. 2021. The global effectiveness of fact-checking: Evidence from simultaneous experiments in argentina, nigeria, south africa, and the united kingdom. *PNAS*.

[Pröllochs2021] Pröllochs, N. 2021. Community-based fact-checking on twitter's birdwatch platform. *arXiv:2104.07175*.

[Resende et al.2019] Resende, G.; Melo, P.; Sousa, H.; Messias, J.; Vasconcelos, M.; Almeida, J.; and Benevenuto, F. 2019. (mis) information dissemination in WhatsApp: Gathering, analyzing and countermeasures. In *WWW*, 818–828. ACM.

[Resnick and others2021] Resnick, P., et al. 2021. Informed crowds can effectively identify misinformation.

[Rossini et al.2021] Rossini, P.; Stromer-Galley, J.; Baptista; et al. 2021. Dysfunctional information sharing on whatsapp and facebook: The role of political talk, cross-cutting exposure and social corrections. *New Media & Society* 23(8).

[Rossini2023] Rossini, P. 2023. Farewell to big data? studying misinformation in mobile messaging applications. *PC*.

[Varanasi, Pal, and Vashistha2022] Varanasi, R. A.; Pal, J.; and Vashistha, A. 2022. Accost, accede, or amplify: Attitudes towards covid-19 misinformation on whatsapp in india. In *CHI*, 1–17.

[Vijaykumar et al.2022] Vijaykumar, S.; Rogerson, D. T.; Jin, Y.; and de Oliveira Costa, M. S. 2022. Dynamics of social corrections to peers sharing covid-19 misinformation on whatsapp in brazil. *JAMIA* 29(1).

[Vraga and others2018] Vraga, E. K., et al. 2018. How providing a source corrects health misperceptions across social media platforms.

[Wood and Porter2019] Wood, T., and Porter, E. 2019. The elusive backfire effect: Mass attitudes' steadfast factual adherence.

[Woolley2022] Woolley, S. C. 2022. Testimony of samuel woolley, for the hearing "a growing threat: The impact of disinformation targeted at communities of color.". Online.