# Structural Dynamics of Harmful Content Dissemination on WhatsApp


## Abstract

WhatsApp has transformed communication, providing a cheaper and more dynamic alternative to traditional SMS, especially through its group chat feature, which enables collective discussions on a wide range of topics. This has made WhatsApp an essential platform for social mobilization, particularly during events such as strikes and political campaigns, where rapid exchange of information is crucial. However, this ease of communication has also raised significant concerns about the spread of misinformation, hate speech, and propaganda. In this study, we explore the dynamics of the dissemination of harmful messages within WhatsApp groups, focusing on their structural characteristics. Using a large dataset, collected through data donation and focused on private groups, covering 5,953 groups in India with more than 5,158,879 messages spanning text, images and videos. Using techniques to robustly reconstruct cascades, we track message propagation throughout the dataset. Our results show that misinformation, hate speech, and propaganda tend to have a significantly greater depth and breadth of dissemination compared to normal messages. In addition, these harmful messages are spread primarily through videos and images, highlighting a distinctive dissemination pattern. However, modality alone cannot fully account for the structural differences in the dissemination between harmful and normal messages, suggesting that the distinct content of harmful messages plays a crucial role in amplifying these differences. The findings highlight the critical role of structural characteristics in the spread of these harmful messages, suggesting that strategies targeting structural characteristics of message chains could be crucial in managing the dissemination of such content on private messaging platforms.


## 1 Introduction

WhatsApp is the world's most popular messaging app, with more than 2.78 billion monthly active users in 180 countries, and is especially dominant in countries like India. In fact, more than 400 million users in India rely on WhatsApp as their main means of communication, making it a central part of daily life for many. Despite its vast user base, studying information dissemination on WhatsApp remains challenging due to its end-to-end encryption, which limits access to message content and makes it difficult to monitor and track harmful material such as misinformation, hate speech, and propaganda. Furthermore, WhatsApp's group chat feature allows large-scale targeting of specific audiences, further exacerbating the challenges of controlling the spread of harmful content. Users can easily create groups and quickly disseminate messages within these groups, increasing the reach of these harmful messages.

Recent events have highlighted the severe real-world consequences of widely distributed harmful messages on WhatsApp. Since 2014, there has been a significant increase in mob violence, often triggered by misinformation spread through WhatsApp. For example, in India, false rumors about child kidnappers circulated on WhatsApp, leading to the deaths of more than two dozen innocent people since April that year. In another case, in August 2018, false rumors about child snatchers on the loose prompted an angry mob in Mexico to attack two men, resulting in their grievous-bodily-harm murder. Beyond violence, WhatsApp has also become a powerful tool for political manipulation. In India, where over 600 million people use WhatsApp, both the BJP[1] and the Congress[2] have been accused of spreading false or misleading information to influence the 900 million voters in the country, highlighting the role of the platform in shaping public opinion [17].

Despite the urgency of understanding how harmful messages spread on WhatsApp, tracking their diffusion remains particularly challenging due to the end-to-end encryption of the platform. Unlike public platforms such as X (formerly Twitter) or Facebook, WhatsApp operates as a private communication network where messages are encrypted upon sending and only decrypted upon receipt. This encryption ensures that even WhatsApp itself cannot access the content exchanged between users. Consequently, developing machine learning algorithms to detect the spread of harmful or false information is virtually impossible, as the platform cannot analyze the messages in transit. Without access to message content, conventional content moderation techniques, as seen on public platforms, cannot be applied. This limitation makes it extremely difficult to monitor or control the dissemination of misinformation on WhatsApp. Although this encryption improves user privacy, it creates significant obstacles for tracking and mitigating the spread of harmful content.

Given the impracticality of content-based moderation, WhatsApp has changed its focus toward the structural characteristics of message dissemination to limit the spread of harmful messages. A key strategy implemented by WhatsApp is to restrict the message forwarding functionality. Specifically, each message can be forwarded to a maximum of five groups, and if the message has already been forwarded multiple times, it can only be forwarded to a single group. Furthermore, once a message has been forwarded more than five times, it is marked with the label "forwarded many times." However, it remains an open question whether harmful messages exhibit structural dissemination patterns that are distinct from normal content. This issue is critical, as the effectiveness of WhatsApp's structural restrictions relies on the assumption that harmful messages follow identifiable patterns, which can be targeted by these restrictions.

To address the above problems, we built a large dataset comprising 5,158,879 messages, including text, images, and videos, collected from 5,953 WhatsApp groups, which is a key channel for the mass


*WWW, 28 April-2 May, 2025, Sydney, Australia*
2025. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX


[1]The Bharatiya Janata Party is the ruling party of India and largest in terms of political representation in the Parliament and state legislatures.
[2]The Indian National Congress (INC), colloquially "the Congress", is one of the oldest political parties in India.

dissemination of information on the platform. This dataset was collected through a field study in India, where participants gave their informed consent to share data from the WhatsApp groups that they were comfortable with, ensuring ethical compliance. To track and analyze the spread of messages, we used privacy-preserving hashing techniques to identify multiple instances of the same message, even with slight variations. For images and videos, we used the PDQ hashing [3], and for text messages we used the locality-sensitive hashing (LSH) [5]. These methods allowed us to detect content variations, such as cropping or encoding changes, capturing exact and modified versions of shared media.

Next, we categorized the dataset into four types: misinformation, hate speech, propaganda, and normal messages. Taking into account the data donation-based sampling strategy, we model the message cascades as generalized tree-like structures and estimate key structural parameters, namely breadth $b$ and depth $h$, to understand the dissemination process. To ensure the robustness of our conclusions, we explored two different cascade structures—*influence cascade* and *network cascade*. The influence cascade assumes that we observe the edges through which the information propagated and reconstruct incomplete cascades using the algorithm developed by Gomez-Rodriguez et al. [8]. The network cascade model assumes that we only observe participating nodes, not the propagation edges, and these edges are inferred based on the time sequence. Introducing a $k$-tree model for network cascades, we estimate the breadth and depth parameters for each message category (misinformation, hate speech, propaganda, and normal messages).

Our analysis revealed that harmful messages have significantly greater breadth and depth compared to normal messages. In addition, these harmful messages are spread primarily through videos and images. We also estimated the scale of dissemination for different message types based on group sizes to derive population-level estimates and found that harmful messages reach a significantly larger audience compared to normal messages. These findings emphasize the need for further investigation into the structural aspects of message propagation, as they are key to developing more effective strategies to limit the spread of harmful content on the platform.

To our knowledge, this is the first study to construct a dataset through data donations focusing on private WhatsApp groups, a notoriously difficult area to study due to end-to-end encryption of the platform. By analyzing the spread of harmful content, such as misinformation, hate speech, and propaganda, we derive key structural parameters (breadth and depth) of information cascades using state-of-the-art algorithms. This work provides unique insights into how harmful content propagates more broadly and deeply compared to normal messages.

## 2 Contributions and Related Work

Harmful messages on social networks are becoming a major issue, with a significant body of research focused on public social media platforms such as X (formerly Twitter) and Facebook. Grinberg et al. [9] find that on Twitter, misinformation exposure is highly concentrated, with just 1% of users encountering 80% of false information. Similarly, Allen et al. [1] study vaccine misinformation on Facebook and find that reports implying vaccines are harmful, not being flagged by fact checkers, had a wider reach and significantly reduced people's willingness to get vaccinated. In another study, Goel et al. [7] highlight the role of hatemongers in spreading hateful speech, noting that these individuals tend to cluster together and form stronger connections within social networks, thus amplifying the spread of harmful content. Likewise, Hristakieva et al. [10] indicate that propaganda spreads more effectively on social media when coordinated between communities.

Although much of the existing research has focused on public platforms like Twitter and Facebook, WhatsApp has recently emerged as another powerful tool to spread harmful messages. However, research on WhatsApp is limited due to several constraints, such as end-to-end encryption, which makes data collection and analysis particularly challenging. To address these limitations, most studies have focused on public WhatsApp groups. For example, Garimella and Tyson [4] propose a generalized method for collecting data from the public WhatsApp groups using Selenium scripts to search for publicly available group invite links through Google. Building on this method, Saha et al. [16] conduct the first large-scale analysis of fear speech in public WhatsApp groups discussing politics in India, and find that such messages spread rapidly and are harder to detect due to their low toxicity. Similarly, Resende et al. [13] examine the dissemination of information in political WhatsApp groups in Brazil, focusing on two social events: the national truck drivers strike and the Brazilian presidential campaign. They discover that misinformation, particularly in the form of images, spreads widely between groups and across platforms during these events. However, focusing solely on public groups is insufficient, as it overlooks the private groups where much of WhatsApp's typical usage occurs. In response, this paper constructs a novel dataset that emphasizes private groups, providing a more accurate reflection of how harmful content spreads on WhatsApp.

Accurately identifying harmful messages remains a pressing issue, as conventional machine learning techniques, though effective on public platforms, are often not scalable for detecting harmful content at a large scale on platforms like Twitter or Facebook, and they are not applicable at all on private platforms like WhatsApp due to end-to-end encryption and other privacy barriers. Given these challenges, researchers have increasingly turned their attention to the structural characteristics of message dissemination as an alternative approach to studying harmful content. Analyzing the structural differences between harmful and regular messages has become crucial in this context.

Studies consistently show that harmful messages spread more effectively than regular ones, highlighting clear structural differences. For example, Vosoughi et al. [18] find that false news on Twitter spreads faster, deeper and more broadly than true news. Similarly, Mathew et al. [12] demonstrate that hateful speech travels farther, spreads faster, and reaches a much wider audience than regular messages. Extending these findings, Maarouf et al. [11] analyze retweet cascades of hateful speech on Twitter and discover that hateful content forms cascades that are 3.5 times larger in size and have 1.2 times greater structural virality, defined as the average distance between all pairs of nodes [6], compared to normal content.

In contrast, the structural characteristics of the dissemination of information on WhatsApp are underexplored. To our knowledge, only Caetano et al. [2] analyze attention cascades, which begin

when a user introduces a topic in a message serving as a starting point for the cascade. Other users contribute by replying either to the original message or to subsequent replies, forming a chain of interactions. Caetano et al. [2] find that attention cascades involving false information in WhatsApp political groups tend to be deeper, reach more users, and last longer than those in non-political groups. However, attention cascades only account for information dissemination within a single group, overlooking cross-group transmission. This cross-group transmission is often crucial for large-scale viral dissemination, where information spreads from one group to another, amplifying its reach. To address this limitation, this paper focuses on cross-group cascades and compares the structural differences between various types of cascades.

However, cascades are complex dynamic entities, and reconstructing the process of cross-group message dissemination is a significant technical challenge in the study of WhatsApp cascades. Several factors limit the reconstruction of cascades. First, it is impossible for anyone, including WhatsApp itself, to fully track the entire message transmission process. As a result, we rely on sampling methods, where specific groups are chosen, and by detecting when messages appear in these groups, we attempt to infer the structural characteristics of the complete information cascade. We adopt a two-pronged approach: First, we reconstruct the transmission process between the observed and sampled group nodes. Second, after obtaining an incomplete, partially observed cascade, we estimate the structural parameters of the complete cascade. To address the first problem, we use two different methods to construct two types of cascades: the **influence cascade** and the **network cascade**. For the influence cascade, we use the algorithm proposed by Gomez-Rodriguez et al. [8], which identifies the optimal network and diffusion process to best explain the observed timings of message appearances in some groups, accounting for external effects and missing nodes, making it highly suitable for our study. For the network cascade, a directed edge is drawn between the nodes $t$ and $s$ if $t$ performed the action before $s$. This allows us to bypass the need for additional algorithms to infer the transmission process between sampled nodes. However, this method sacrifices some interpretability because it does not directly illustrate the message transmission process as clearly as the influence cascade does. In this study, the network cascade serves primarily to ensure the robustness of our conclusions by relying solely on temporal ordering of the nodes without inferring detailed diffusion paths. For the second problem, we employ the algorithm proposed by Sadikov et al. [15]. The core idea is to calculate the expected properties of the incomplete cascade, such as the number of nodes and edges, based on the theoretical k-tree model and the sampling probability (details in Section 4.1). By minimizing the differences between these theoretical values and the actual observed data, the algorithm accurately estimates the structural parameters of the tree model, even when up to 90% of the data are missing. For completing the **influence cascade**, we combines two different algorithms proposed by Gomez-Rodriguez et al. [8] and Sadikov et al. [15] to estimate the structural parameters of the complete information cascade. To complete the **network cascade**, we only use the second algorithm proposed by Sadikov et al. [15]. Another approach frequently used to reconstruct cascades is based on the Steiner tree problem, where a minimum-cost tree is sought to span all reported nodes while preserving the order of observed timestamps [14, 19]. However, this method is unsuitable for our scenario because of its limitations in handling large amounts of missing data, which is a feature of our dataset explained next.

## 3 Dataset

We collected a dataset of WhatsApp group messages through data donations from 3,500 users in the northern Indian state of Uttar Pradesh, corresponding to 5,953 WhatsApp groups. The data collection spanned from October 2023 to June 2024, yielding more than 5 million messages. To ensure a representative sample of villages and capture diverse demographics, we carefully designed our sampling method.

**Sampling Procedure**. Our sampling procedure involved randomly selecting 10 districts within Uttar Pradesh and then randomly picking 10 villages from each chosen district. For each selected village, we obtained census data to establish the baseline distribution of the population in terms of age, caste, and religion.[3] Based on this distribution, our survey team visited each village and sought the consent of the participants until we filled the quotas corresponding to the age, caste, and religion demographics. We opted for a quota sampling method instead of a purely random sample due to practical considerations related to the uptake of our data donation process.

**Data Donation Process**. The on-the-ground protocol involved surveyors reaching out to participants to obtain informed consent and explain our data collection and anonymization protocols. We employed a custom-built data donation tool to facilitate users in donating their WhatsApp group data. Only groups with more than five participants and a certain level of activity were eligible for donation. Upon completion of the donation process, our tool collected the following data: (i) All messages from the two months preceding and the two months following the date of donation. (ii) Anonymized contacts from the user's phone contacts; and, (iii) Anonymized group membership information.

**Dataset Statistics**. The collected dataset comprises over 5 million messages from more than 5,900 WhatsApp groups during October 2023 to June 2024. The median group size was 104 members, indicating that most groups were large and engaged in discussions around political and religious identity, caste, region, and related topics.

**Annotations**. We annotated a subset of messages to identify content that contains misinformation, hate speech, and political propaganda. Specifically, we annotated all 2,019 pieces of content that were marked as "forwarded many times" and shared during October to December 2023.[4] The annotations were performed by a professional fact-checker who was well-versed with the content and cultural context of the data. Out of the annotated messages, we identified 401 instances of misinformation, 111 instances of hate speech, and 116 instances of propaganda.[5]

[3]Due to practical limitations during our data donation pilot, we did not attempt to obtain a sample representative of gender.

[4]We used definitions of misinformation, hate speech, and propaganda from the Facebook Community Standards documents (https://transparency.meta.com/policies/community-standards/misinformation).

[5]This sample of misinformation, hate and propaganda is not an ideal sample since it does not cover all the time period of our messages. However, an annotation of thousands of pieces of content is very tedious and time consuming, requires experts,

**Ethical Considerations**. The data collection process was approved by the Institutional Review Board (IRB) at multiple participating institutions (details anonymized). We took extreme care to minimize the amount of data collected and to protect personal information. Personal identifiers, including phone numbers, emails, and faces in images, were anonymized before storage on our servers. Importantly, the data analyzed in this study did not include the content of the messages; only metadata were analyzed to ensure the privacy of participants.

## 4 Methodology

In this paper, we focus on the structural characteristics of information dissemination, particularly the breadth and depth of cascades. We assume that the information propagates in a tree-like structure and attempt to reconstruct the dissemination process based on the available data. A key challenge lies in the fact that our dataset does not allow us to fully trace the entire information transmission process. This limitation is almost inevitable when analyzing platforms such as WhatsApp, where even the platform itself is unable to track the entire path of information dissemination. Even if we assume that the platform can trace all instances of the same content, the information could originate from two distinct transmission paths, with each source independently influenced by external factors.

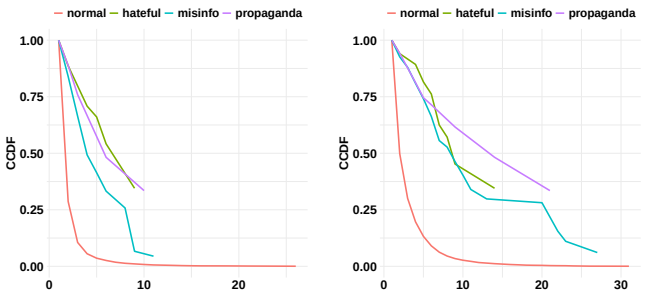
As such, accurately reconstructing the entire cascade of information transmission is almost impossible. Instead, we focus on providing reliable estimates for certain structural characteristics of the cascades with missing data. Another challenge is that in our dataset we can only track the time at which information reaches nodes, but we do not know whether there are direct links between nodes. To overcome this issue, we compared two different approaches. The first approach involves estimating the probability that a directed edge exists between two nodes based on the time information reaches each node. This allows us to reconstruct an incomplete *influence cascade*, from which we then estimate the structural characteristics of the complete influence cascade. The second approach involves the construction of a *network cascade*, where a directed edge is drawn between nodes $t$ and $s$ if $t$ has performed the action before $s$. Then, on the basis of the incomplete network cascade, we estimate the structural characteristics of the complete network cascade.

Ultimately, our goal is to draw reliable conclusions about the structural features of information dissemination based on incomplete cascade data and to later understand the population-level differences (e.g., how many people reached) between harmful messages, such as misinformation, hate speech and propaganda, and normal information.

### 4.1 Influence Cascade Reconstruction

The *influence cascade* model focuses on influence relationships within an action sequence. For example, an action sequence can begin with an active node, followed by one of its neighbors, neighbors of neighbors, etc. For the influence cascade, the primary challenge is the lack of observable influence relationships between nodes, since our dataset only provides information on when a message reaches a group, without details on the diffusion process between groups. To overcome this, we employ the methods of Gomez-Rodriguez et al. [8] to reconstruct *complete influence cascades* from partial data. The algorithm works by inferring the information diffusion process from the observed times when nodes (or groups) adopt information. A key assumption is that if information appears in two groups at closely related times, it is more likely that a directed edge exists between them. Using this temporal proximity, the algorithm identifies potential influence cascades. In addition, the algorithm accounts for external effects and the presence of missing nodes, making it suitable for reconstructing incomplete influence cascades more accurately. To effectively apply this algorithm, if a message appears multiple times within a group, we only consider the timestamp of its first occurrence. Moreover, we excluded messages that appeared only in a single group from the dataset, as such messages do not contribute to the cascade structure. This ensures that the cascade model reflects meaningful propagation between multiple groups.

Although this reconstruction is based on incomplete data, it still allows us to examine key structural properties such as **maximum breadth** and **depth**. These properties differ from the parameters $b$ and $h$ used in the tree model, which are also referred to as breadth and depth, but represent different aspects of the cascade. Specifically, depth in this context is defined as the total number of edges from the root node to the leaf node along the longest path in the incomplete influence cascade, while breadth refers to the number of messages at a particular depth level in the cascade. Our focus is on the maximum breadth, defined as the largest number of messages at any depth level. It is worth noting that the terms depth and breadth used here differ slightly from the depth and breadth described in the tree model later in the paper, and care should be taken to distinguish between these two contexts to avoid potential confusion.




**(a) Breadth** **(b) Depth**

**Figure 1: Comparison of CCDF for Breadth and Depth**

By analyzing the maximum breadth and depth for each message type, we aim to demonstrate that even in the sampled dataset, harmful messages and normal messages exhibit significant differences in their propagation patterns. To support this, we present complementary cumulative distribution function (CCDF) plots to compare how different message types spread across groups. As shown in Figures 1a and 1b, harmful messages demonstrate significantly greater breadth and depth compared to normal message types, indicating their broader and deeper propagation between groups. In addition,

and does not scale well. We are in the process of collecting expert annotations for the rest of our dataset and will update the manuscript with numbers that compare misinformation on the entire dataset.

propaganda messages exhibit greater breadth and depth compared to misinformation. These findings highlight that, despite the limitations of the sampled data, the observed structural differences are robust and consistent.

Using this partially observed cascade dataset, the next step is to develop an algorithm that can infer the properties of the entire cascade tree within this network. In particular, given only a fraction $C'$ of the complete cascade $C$, our goal is to estimate key properties of the complete cascade, such as breadth, depth, or size, that is, the total number of people exposed to the message. To accomplish this, we use the method developed by [15]. We begin by assuming that the complete information diffusion process follows a tree model. Each node in the sampled cascade tree is included with a probability of $p$. If the parameters of the tree model are known, then we can estimate the properties of the complete cascade $C$, based on the properties derived from the tree model. The parameters of the tree model are determined by matching the theoretical values calculated from the tree model with the measured values on the sampled cascade $C'$. Specifically, a sampled tree, denoted as $\Gamma(p, b, h)$, is generated with a depth $h$, a branching factor or breadth $b$, and a sampling probability $p$. From this sampled tree, we can derive theoretical expressions for certain properties that we then matched to the observed data to estimate the values of the parameters $b$ and $h$. For example, the expected number of nodes in the sampled tree $\Gamma(p, b, h)$ is $p\frac{b^{h+1}-1}{b-1}$, which is matched to the number of nodes observed in the reconstructed cascades. For a list of all the matching properties and detailed proofs, see [15].

In our case, the parameter $p$ is the sampling probability of each group, which we can calculate based on the data collection process (as described in Section 3). Practically, we set the value of $p$ to 0.02 (2%) given that we were sampling roughly 1% of a village and most of the users provided almost 90% of the groups they had. Then, we can estimate the breadth and depth by minimizing the sum of errors between the values of these properties from our dataset and those calculated from the theoretical sampled tree model $\Gamma(p, b, h)$.

## 4.2 Network Cascade Reconstruction

To further validate the reasonableness of the results of the influence cascade reconstruction, we consider the *network cascade* model. In contrast to the influence cascade model, the *network cascades* are built on the scenario where only the action of a node $s$ is known, without clarity on who influenced that action. In this case, a directed edge is drawn between nodes $t$ and $s$ if $t$ performs the action before $s$, which means that no algorithm is required to reconstruct diffusion processes between the sampled groups. However, the sampled tree model $\Gamma(p, b, h)$ is not appropriate for *network cascades*, as nodes may now have more than one incoming edge. Consequently, the network cascades form a directed acyclic graph (DAG) rather than a tree. To address this issue, we need to replace the tree model $\Gamma(p, b, h)$ with a $k$-tree model $\Gamma(p, b, k, h)$. To convert the tree model into a $k$-tree model, each node is supplemented with $k-1$ extra edges, linking it to its $k-1$ nearest ancestors, beginning with its grandparent. Figure 2 illustrates the tree model, the influence cascade, the $k$-tree model, and the network cascade. Using the $k$-tree model, we can estimate the three parameters $b$, $h$, and $k$ by minimizing the sum of errors between the observed values of specific properties in our dataset and those computed from the theoretical $k$-tree model $\Gamma(p, b, k, h)$.

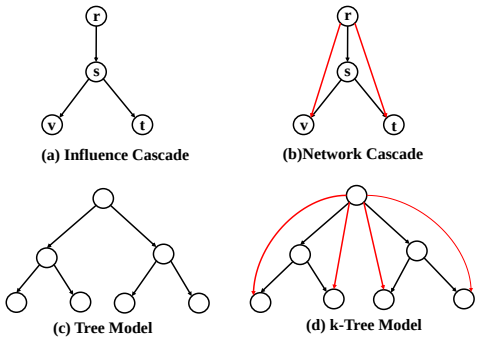



**Figure 2: (a) Influence cascade. (b) Network cascade. (c) Tree model with breadth $b = 2$ and depth $h = 2$. (d) K-tree model with breadth $b = 2$, number of parents $k = 2$ depth $h = 2$.**

In Figure 2(a), for the influence cascade, the message first appears in group $r$, and then someone from group $r$ forwards it to group $s$. Subsequently, someone from group $v$ forwards the message to both group $v$ and group $t$. However, in Figure 2(b), we only know the time at which the message appears in each group. In this case, we know that $t_r < t_s < t_v = t_t$, so we connect the nodes in temporal order, forming a network cascade. Compared to the influence network, the red directed edges edge$(r, v)$ and edge$(r, t)$ are now spurious. As a result, the network cascade becomes a DAG.

## 4.3 Population-Level Impact Estimation

Based on the estimates of the influence cascade and the network cascade, we can derive the parameters required for both the complete tree model and the $k$-tree model. This allows us to understand the structural characteristics of specific types of information transmission, such as whether misinformation has greater breadth or depth compared to normal messages. Additionally, we can calculate their respective reach, such as how many groups an average misinformation message is expected to propagate through. By analyzing our dataset, we can also estimate the distribution of group sizes. This enables us to combine the number of groups traversed by a complete information cascade with the size of those groups, providing insights into the population-level impact of specific information types, namely, how many people are ultimately affected by different types of message.

# 5 Results

In this analysis, our objective is to understand the dissemination patterns of different message types, including harmful messages such as misinformation, hate speech, and propaganda, as well as normal messages. We start by analyzing the structural characteristics of message dissemination through two types of cascades: the influence cascade and the network cascade. In addition, we investigate the role of message modality, whether chat, image, or video, in influencing dissemination patterns. Finally, we provide population-level estimates to quantify the total number of people typically

affected by each message type, combining group size data with the average number of groups through which a message passes. This comprehensive approach allows us to identify structural differences in the spread of harmful content, offering key insights into their broader societal impact.

## 5.1 Influence Cascade

Based on insights into structural characteristic differences from incomplete influence cascades, we can extend the analysis to examine the structural characteristics of complete influence cascades across four message types: misinformation, hate speech, propaganda, and normal messages. Specifically, we model the complete influence cascade using a tree structure, characterized by two key parameters: the branching factor, $b$ (breadth), which indicates the average number of connections a node generates, and the depth, $h$, which represents the maximum number of layers in the tree, or how deep the information travels. We assume that the incomplete influence cascade observed in our dataset is a sampled version of the complete cascade, with some connections or nodes missing due to the limitations of the data. It is important to note that, when labeling messages as harmful or not, privacy restrictions prevented us from classifying all messages in the dataset. As a result, we focused only on messages that were forwarded many times, specifically those with a forwarding score of five or more. We applied our harmful message classification to this subset of messages. Therefore, to ensure the robustness of our findings, we also further categorized normal messages. We extracted normal messages with a forwarding score greater than or equal to five and assessed whether these messages exhibit significant structural differences compared to harmful messages. This extended classification allows us to draw more general conclusions regarding the structural distinctions between harmful and normal messages within the context of complete influence cascades.

**Table 1: Mean and Standard Deviation of Parameters by Content Type**

| Content Type | $\mu_b$ | $\sigma_b$ | $\mu_h$ | $\sigma_h$ |
|---|---|---|---|---|
| Hateful | 3.78 | 0.575 | 4.89 | 0.244 |
| Misinformation | 3.68 | 0.579 | 4.86 | 0.261 |
| Propaganda | 3.82 | 0.660 | 4.92 | 0.298 |
| Normal | 2.85 | 0.388 | 4.50 | 0.167 |
| Normal ($F.S. \geq 5$) | 3.37 | 0.546 | 4.72 | 0.235 |

For each type of message, we estimate the values of $b$ and $h$ to capture the spread dynamics within the influence cascade. The results, shown in Table 1, indicate that harmful messages, including misinformation, hate speech, and propaganda, exhibit higher values of $b$ and $h$ compared to normal messages. This suggests that harmful content spreads more broadly and deeply, reaching a wider audience and traveling further into the network.

Specifically, while all harmful messages show higher breadth and depth compared to normal messages and normal messages with a forwarding score greater than or equal to five, there are notable differences among the harmful message types. Propaganda has the highest values for both breadth ($b = 3.82$) and depth ($h = 4.92$), followed closely by hate speech with a breadth of $b = 3.78$ and depth of $h = 4.89$. Misinformation, though still significantly larger than normal messages, has slightly lower values ($b = 3.68$ and $h = 4.86$). This suggests that, although misinformation spreads widely, propaganda and hate speech tend to spread even more broadly and deeply on WhatsApp.

These findings highlight the distinctive structural characteristics of harmful messages in terms of their influence cascades, providing crucial insights into how these messages disseminate through WhatsApp's network.

## 5.2 Network Cascade

To ensure the robustness of our results, we also analyzed the fit of the network cascade model. Unlike the influence cascade model, where we infer the edges between nodes, in the network cascade, we connect nodes solely based on the order in which they received the information. This approach can result in nodes having multiple incoming edges, transforming the cascade into a DAG rather than a tree.

To address this complexity, we introduced the $k$-tree model. In addition to estimating the parameters $b$ (breadth) and $h$ (depth), we also incorporated the parameter $k$, which accounts for the multiple incoming edges in the DAG structure. Specifically, each node is supplemented with $k - 1$ additional edges, linking it to its $k - 1$ nearest ancestors, starting with its grandparent. This extension allows us to better capture the structure of the network cascade.

**Table 2: Mean and Standard Deviation of Parameters by Content Type**

| Content Type | $\mu_b$ | $\sigma_b$ | $\mu_h$ | $\sigma_h$ | $\mu_k$ | $\sigma_k$ |
|---|---|---|---|---|---|---|
| Hateful | 3.46 | 0.711 | 5.29 | 0.520 | 1.00 | 0.000 |
| Misinformation | 3.21 | 0.383 | 5.18 | 0.427 | 1.15 | 0.315 |
| Propaganda | 3.58 | 0.684 | 5.17 | 0.349 | 1.00 | 0.000 |
| Normal | 2.77 | 0.259 | 4.58 | 0.348 | 1.00 | 0.050 |
| Normal ($F.S. \geq 5$) | 3.08 | 0.423 | 5.06 | 0.595 | 1.01 | 0.088 |

Our results, detailed in Table 2, show that the network cascade analysis aligns with the trends observed in the influence cascade. Harmful messages, including hateful, misinformation, and propaganda, exhibit larger values of breadth and depth compared to normal messages.

Specifically, hateful and propaganda messages show larger breadth values ($b = 3.46$ and $b = 3.58$, respectively) compared to misinformation ($b = 3.21$), indicating that hateful and propaganda content tends to spread to a broader range of groups. These variations highlight distinct patterns in the way different types of harmful messages are disseminated across the network, further confirming the robustness of these structural characteristics in the dissemination of harmful messages.

Furthermore, the mean value of $k$ remains close to 1 in the table, which further validates the stability of the influence cascade model. The introduction of the parameter $k$ in the network cascade appears to be redundant, as the structure often did not form a DAG but instead continued to follow a tree-like transmission pattern.

To further ensure the robustness of our results, we performed Wilcoxon rank-sum tests on the breadth and depth of the influence cascade and network cascade for the four different message types. The results of the Wilcoxon rank-sum tests on the breadth and depth of both the influence cascade and the network cascade for the four message types are provided in the appendix. These results, shown in Tables 4 and 5, confirm that the structural differences in message dissemination between harmful and normal messages are statistically significant in both cascades. Additionally, further analysis reveals significant differences between propaganda and misinformation, with propaganda demonstrating notably greater breadth and depth.

## 5.3 The Impact of Message Modality on Dissemination Patterns

In this section, we explore the reasons behind the broader dissemination of harmful messages—such as hateful content, misinformation, and propaganda—compared to normal messages. Based on our analysis of the dataset, Figure 3 highlights significant variations in the distribution of dissemination modalities (chat, image, video) across different content types. For example, video is the dominant modality for both hateful speech and political propaganda, comprising 88.9% and 87.6% of their respective message distributions. In contrast, normal content exhibits a more balanced distribution, with chat accounting for 12.4%, image for 51%, and video for 36.6%. These differences suggest that hateful speech and political propaganda rely heavily on video dissemination, while misinformation is more evenly split between image and video, with a smaller emphasis on video. This distribution pattern suggests that the modality through which messages are shared may influence the structure of their transmission, potentially contributing to the wider reach observed in harmful messages.

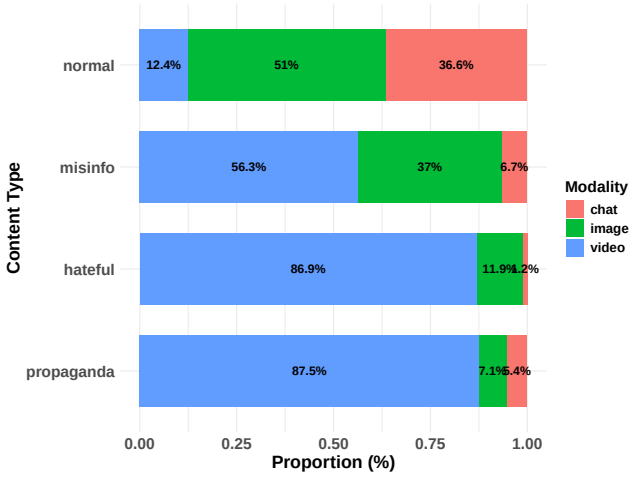



**Figure 3: Proportion of different modality types within each content type.**

To test this hypothesis, we classified messages by modality (chat, image, video) and analyzed the breadth and depth of both influence and network cascades for each modality, as shown in Table 3. Our results indicate that video messages consistently exhibit a greater breadth of dissemination compared to chat and image messages in both cascade models. However, despite these differences, the structural characteristics (breadth, depth, and complexity) between chat, video, and image do not vary significantly. This suggests that while modality may contribute to some differences in dissemination, it alone is not sufficient to explain why harmful messages, such as propaganda and hateful speech, spread more widely and deeply than normal messages.

**Table 3: Comparison of Influence Cascade and Network Cascade for different modalities**

| | Modality | $\mu_b$ | $\sigma_b$ | $\mu_h$ | $\sigma_h$ | $\mu_k$ | $\sigma_k$ |
|---|---|---|---|---|---|---|---|
| **Network** | chat | 2.76 | 0.258 | 4.57 | 0.33 | 1.00 | 0.067 |
| | video | 2.88 | 0.351 | 4.78 | 0.535 | 1.01 | 0.071 |
| | image | 2.75 | 0.228 | 4.55 | 0.284 | 1.00 | 0.030 |
| **Influence** | chat | 2.84 | 0.389 | 4.49 | 0.167 | - | - |
| | video | 3.06 | 0.524 | 4.59 | 0.225 | - | - |
| | image | 2.81 | 0.332 | 4.48 | 0.142 | - | - |

Therefore, the broader dissemination of harmful messages likely results from a combination of factors beyond just modality, indicating the need for further research to explore deeper drivers of dissemination dynamics in harmful content.

## 5.4 Population-Level Estimation

In this final section, we estimate the population-level impact of different message types by determining how many people are typically affected. From our dataset, we know the number of participants in each WhatsApp group, allowing us to estimate the average group size. Combined with the results from both the influence cascade and network cascade analyses, where we can calculate the average number of groups each type of message passes through, we can approximate the total number of individuals exposed to each type of message.

Figure 4 presents the distribution of the group sizes on a logarithmic scale. For our dataset, the mean group size is 81.25. By multiplying the estimated average group size by the number of groups that a message typically reaches based on both the influence cascade and the network cascade, we can approximate the population-level impact for each type of message. This approach allows us to quantify the broader societal impact of harmful messages, such as misinformation, hate speech, and propaganda, compared to normal messages. In Figure 5, we plot the distribution of population-level estimates based on influence cascade parameters, showing the number of individuals affected by different types of messages. The corresponding figure for the network cascade is included in the appendix. Our findings suggest that harmful messages, due to their larger breadth and depth, tend to reach significantly more people than normal messages. According to our estimates, harmful messages affect approximately five times more individuals than normal messages. Even when focusing on normal messages with a forwarding score greater than or equal to five, harmful messages still reach

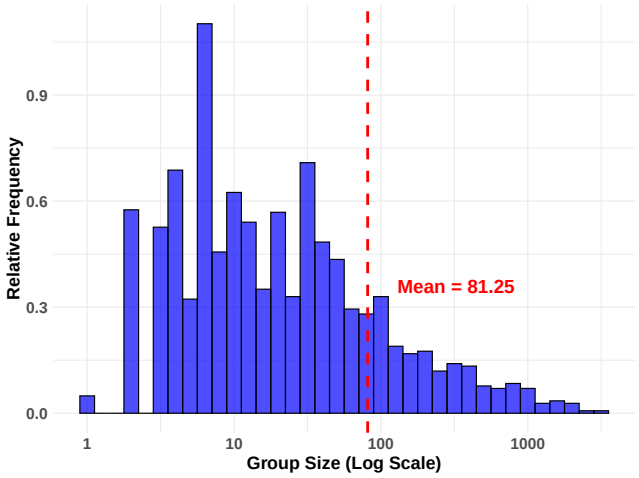



**Figure 4: Distribution of group sizes on a log scale. The red dashed line represents the mean group size.**

nearly twice the audience compared to these highly forwarded normal messages. At the same time, there are noticeable differences among harmful messages. Propaganda and hateful speech tend to reach a larger audience, while the impact of misinformation is relatively smaller.

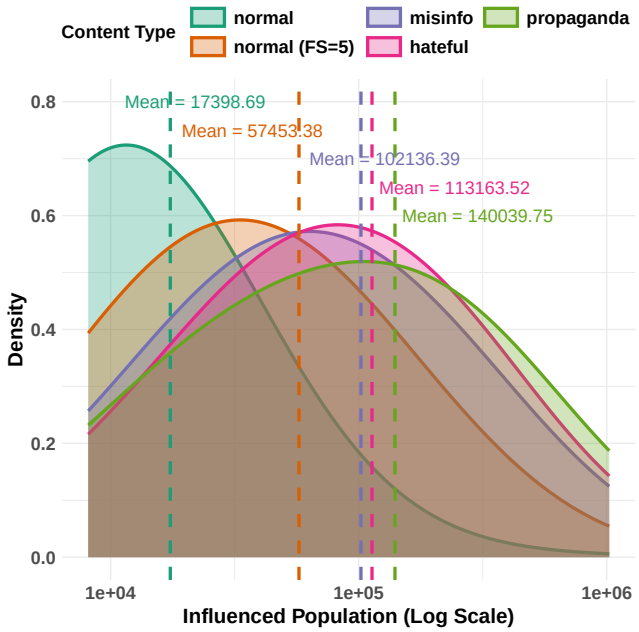



**Figure 5: Density of influenced population by content type based on influence cascade. Vertical dashed lines indicate the mean influenced population for each content type.**

## 5.5 Demographic Analysis of Cascade Structures

In this section, we build on the conclusions of this study by incorporating demographic features to further investigate the structural properties of cascades. For each cascade, we analyzed the demographic characteristics of the individual who first initiated the message, with the aim of determining whether there are structural differences in cascades started by people with varying demographic backgrounds. Specifically, we considered five demographic features: caste, religion, income, education, and age. Detailed results are provided in the appendix.

For most of the demographic features, we did not find any significant structural differences in the cascades. Although this suggests that demographic factors may not have a substantial impact on the breadth or depth of information cascades, further analysis is required to draw definitive conclusions. Future studies could explore these factors more thoroughly to understand whether subtle or context-specific effects could exist.

## 6 Discussion

In this paper, we study the structural dynamics of harmful content dissemination in WhatsApp groups. We construct a new dataset through data donations, covering 5,953 groups in India and consisting of 5,158,879 messages that span text, images, and videos. Using this large-scale dataset, we apply an algorithm developed by Gomez-Rodriguez et al. [8] to reconstruct the dissemination paths between the groups in our sampled data. We then use the algorithm developed by Sadikov et al. [15] to estimate two key structural parameters of the complete information diffusion process from partially observed cascades in our sample: breadth and depth. Our study reveals several key findings: First, harmful messages, such as misinformation, hate speech, and propaganda, tend to have significantly greater breadth and depth of dissemination compared to normal messages. Furthermore, propaganda exhibits a wider and deeper spread compared to misinformation. We also found that the dissemination of harmful messages is primarily driven through video and image formats. However, differences in message modality alone are insufficient to fully explain the significant structural differences in the dissemination processes between harmful and normal messages. Finally, we estimate the population-level impact of harmful messages, finding that, on average, harmful messages affect approximately five times more people than normal messages.

To our knowledge, this is the first effort to build a new dataset on private groups based on data donation. However, we also acknowledge certain limitations in the methodology used in this paper. First, the second algorithm developed by [15] is based on several strong assumptions, such as uniform sampling. In future work, our aim is to extend this model to obtain more accurate population-level estimates. In addition, we intend to further investigate the underlying factors that contribute to the structural differences between harmful and normal messages. Finally, we hope to integrate demographic data to explore how harmful messages affect different demographic groups differently.

## References


[1] Jennifer Allen, Duncan J. Watts, and David G. Rand. 2024. Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. *Science* 384, 6699 (2024), eadk3451. https://doi.org/10.1126/science.adk3451

[2] Josemar Alves Caetano, Gabriel Magno, Marcos Gonçalves, Jussara Almeida, Humberto T Marques-Neto, and Virgílio Almeida. 2019. Characterizing attention cascades in whatsapp groups. In *Proceedings of the 10th ACM conference on web science*. 27–36.

[3] Hany Farid. 2021. An overview of perceptual hashing. *Journal of Online Trust and Safety* 1, 1 (2021).

[4] Kiran Garimella and Gareth Tyson. 2018. Whatapp doc? a first look at whatsapp public group data. In *Proceedings of the international AAAI conference on web and social media*, Vol. 12.

[5] Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. 1999. Similarity search in high dimensions via hashing. In *Vldb*, Vol. 99. 518–529.

[6] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J. Watts. 2016. The Structural Virality of Online Diffusion. *Management Science* 62, 1 (2016), 180–196.

[7] Vasu Goel, Dhruv Sahnan, Subhabrata Dutta, Anil Bandhakavi, and Tanmoy Chakraborty. 2023. Hatemongers ride on echo chambers to escalate hate speech diffusion. *PNAS Nexus* 2, 3 (02 2023), pgad041. https://doi.org/10.1093/pnasnexus/pgad041

[8] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. 2012. Inferring Networks of Diffusion and Influence. *ACM Trans. Knowl. Discov. Data* 5, 4, Article 21 (feb 2012), 37 pages. https://doi.org/10.1145/2086737.2086741

[9] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363, 6425 (2019), 374–378. https://doi.org/10.1126/science.aau2706

[10] Kristina Hristakieva, Stefano Cresci, Giovanni Da San Martino, Mauro Conti, and Preslav Nakov. 2022. The Spread of Propaganda by Coordinated Communities on Social Media. In *Proceedings of the 14th ACM Web Science Conference 2022* (Barcelona, Spain) *(WebSci '22)*. Association for Computing Machinery, New York, NY, USA, 191–201. https://doi.org/10.1145/3501247.3531543

[11] Abdurahman Maarouf, Nicolas Pröllochs, and Stefan Feuerriegel. 2024. The Virality of Hate Speech on Social Media. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 186 (April 2024), 22 pages. https://doi.org/10.1145/3641025

[12] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of Hate Speech in Online Social Media. In *Proceedings of the 10th ACM Conference on Web Science* (Boston, Massachusetts, USA) *(WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/3292522.3326034

[13] Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. 2019. (Mis) information dissemination in WhatsApp: Gathering, analyzing and countermeasures. In *The World Wide Web Conference*. 818–828.

[14] Polina Rozenshtein, Aristides Gionis, B Aditya Prakash, and Jilles Vreeken. 2016. Reconstructing an epidemic over time. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1835–1844.

[15] Eldar Sadikov, Montserrat Medina, Jure Leskovec, and Hector Garcia-Molina. 2011. Correcting for missing data in information cascades. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (Hong Kong, China) *(WSDM '11)*. Association for Computing Machinery, New York, NY, USA, 55–64. https://doi.org/10.1145/1935826.1935844

[16] Punyajoy Saha, Binny Mathew, Kiran Garimella, and Animesh Mukherjee. 2021. "Short is the Road that Leads from Fear to Hate": Fear Speech in Indian WhatsApp Groups. In *Proceedings of the Web conference 2021*. 1110–1121.

[17] Hindustan Times. 2019. For PM Modi's 2019 campaign, BJP readies its WhatsApp plan — hindustantimes.com. https://www.hindustantimes.com/india-news/bjp-plans-a-whatsapp-campaign-for-2019-lok-sabha-election/story-lHQBYbxwXHaChc7Akk6hcI.html. [Accessed 01-03-2024].

[18] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. https://doi.org/10.1126/science.aap9559

[19] Han Xiao, Polina Rozenshtein, Nikolaj Tatti, and Aristides Gionis. 2018. Reconstructing a cascade from temporal observations. In *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 666–674.


## A Wilcoxon Rank-Sum Test Results for the Breadth and Depth of Influence and Network Cascades

To ensure the robustness of our results, we performed one-sided Wilcoxon rank sum tests to compare the breadth and depth of cascades across different types of content for influence and network cascades, such as testing hypotheses whether the depth of misinformation is significantly greater than that of normal messages, as shown in Tables 4 and 5. The results demonstrate that, for both the influence and network cascades, there are significant structural differences between harmful content types (such as misinformation, hate speech, and propaganda) and normal messages.

For the breadth of cascades, the comparisons between misinformation, hate speech, propaganda, and normal messages consistently show highly significant differences (p-value < 1e-5). Furthermore, pairwise comparisons between propaganda and misinformation reveal notable differences, indicating that even among harmful messages, the structural properties of dissemination vary significantly.

For the depth of cascades, similar patterns emerge. Harmful messages such as misinformation, hate speech, and propaganda exhibit significantly deeper cascades compared to normal messages, with p-values < 1e-5 in most comparisons. This indicates that harmful messages penetrate the network more deeply than regular content. In addition, the difference between propaganda and misinformation remains significant, further emphasizing the unique structural dynamics between different types of harmful content.

These results reinforce the conclusion that harmful messages spread not only more broadly, but also more deeply through networks compared to normal messages. Furthermore, the differences between various types of harmful content suggest that certain types of harmful content, such as propaganda, may be more effective at reaching a broader and deeper audience than others, such as misinformation. The significance of these results across both influence and network cascades ensures the robustness of our findings, confirming the reliability of our cascade reconstruction methods in capturing the structural differences in message dissemination.

**Table 4: Wilcoxon Rank-Sum Test Results for the Breadth of Influence and Network Cascades**

| | Content Type Comparison | Test Statistic | p-value |
|---|---|---|---|
| **Influence** | Misinfo - Normal | 731963991.00 | < 1e-5 |
| | Hateful - Normal | 100649135.00 | < 1e-5 |
| | Propa - Normal | 132513655.00 | < 1e-5 |
| | Misinfo - Normal ($F.S. \geq 5$) | 48347735.00 | < 1e-5 |
| | Hateful - Normal ($F.S. \geq 5$) | 6953780.00 | < 1e-5 |
| | Propa - Normal ($F.S. \geq 5$) | 9256863.00 | < 1e-5 |
| | Hateful - Misinfo | 114896.50 | 0.0215 |
| | Propa - Misinfo | 158536.00 | 0.0006 |
| | Propa - Hateful | 20567.50 | 0.0563 |
| **Network** | Misinfo - Normal | 752201958.00 | < 1e-5 |
| | Hateful - Normal | 97630490.50 | < 1e-5 |
| | Propa - Normal | 136109337.50 | < 1e-5 |
| | Misinfo - Normal ($F.S. \geq 5$) | 47824077.00 | < 1e-5 |
| | Hateful - Normal ($F.S. \geq 5$) | 6336116.00 | < 1e-5 |
| | Propa - Normal ($F.S. \geq 5$) | 9144308.00 | < 1e-5 |
| | Hateful - Misinfo | 111025.50 | 0.1064 |
| | Propa- Misinfo | 169128.00 | < 1e-5 |
| | Propa - Hateful | 22290.00 | 0.0008 |

**Table 5: Wilcoxon Rank-Sum Test Results for the Depth of Influence and Network Cascades**

| | Content Type Comparison | Test Statistic | p-value |
|---|---|---|---|
| **Influence** | Misinfo - Normal | 731681544.00 | < 1e-5 |
| | Hateful - Normal | 100716850.00 | < 1e-5 |
| | Propaganda - Normal | 132021521.50 | < 1e-5 |
| | Misinfo - Normal ($F.S. \geq 5$) | 48331757.00 | < 1e-5 |
| | Hateful - Normal ($F.S. \geq 5$) | 6935536.00 | < 1e-5 |
| | Propa- Normal ($F.S. \geq 5$) | 9219857.50 | < 1e-5 |
| | Hateful - Misinfo | 107938.50 | 0.2661 |
| | Propaganda - Misinfo | 159914.00 | 0.0002 |
| | Propaganda - Hateful | 20582.50 | 0.0548 |
| **Network** | Misinfo - Normal | 760922645.00 | < 1e-5 |
| | Hateful - Normal | 104852797.50 | < 1e-5 |
| | Propaganda - Normal | 137310499.50 | < 1e-5 |
| | Misinfo - Normal ($F.S. \geq 5$) | 46107801.00 | < 1e-5 |
| | Hateful - Normal ($F.S. \geq 5$) | 6669863.00 | < 1e-5 |
| | Propa- Normal ($F.S. \geq 5$) | 8381746.50 | < 1e-5 |
| | Hateful - Misinfo | 110349.50 | 0.1335 |
| | Propaganda - Misinfo | 148478.00 | 0.0685 |
| | Propaganda - Hateful | 19495.00 | 0.2692 |

## B Validation of Cascade Reconstruction Method Using Forwarding Scores

In Table 6, we categorize the data based on the forwarding score, a feature of WhatsApp that records the number of times a message is forwarded. When a message is forwarded more than five times, it is labeled "forwarded many times." Using this characteristic, we reclassify the cascade dataset and apply the same reconstruction algorithm to datasets with different forwarding scores. We expect that as the forwarding score increases, the corresponding parameters, breadth ($b$) and depth ($h$), will also increase. This would further validate the accuracy of our method.

**Table 6: Comparison Between Different Forwarding Scores for Influence Cascade and Network Cascade**

| | F.S. | $\mu_b$ | $\sigma_b$ | $\mu_h$ | $\sigma_h$ | $\mu_k$ | $\sigma_k$ |
|---|---|---|---|---|---|---|---|
| **Network** | 0 | 2.73 | 0.222 | 4.53 | 0.274 | 1.00 | 0.0393 |
| | 1 | 2.73 | 0.194 | 4.52 | 0.237 | 1.00 | 0.0396 |
| | 2 | 2.77 | 0.208 | 4.57 | 0.261 | 1.00 | 0.0000 |
| | 3 | 2.82 | 0.252 | 4.64 | 0.358 | 1.01 | 0.1080 |
| | 4 | 2.88 | 0.299 | 4.76 | 0.516 | 1.01 | 0.1070 |
| | $\geq 5$ | 3.08 | 0.426 | 5.06 | 0.591 | 1.01 | 0.1000 |
| **Influence** | 0 | 2.79 | 0.326 | 4.47 | 0.140 | - | - |
| | 1 | 2.78 | 0.292 | 4.47 | 0.125 | - | - |
| | 2 | 2.83 | 0.301 | 4.49 | 0.129 | - | - |
| | 3 | 2.92 | 0.421 | 4.53 | 0.180 | - | - |
| | 4 | 3.05 | 0.538 | 4.58 | 0.230 | - | - |
| | $\geq 5$ | 3.38 | 0.550 | 4.73 | 0.237 | - | - |

As shown in Table 6, both the breadth and depth increase as the forwarding score increases. In particular, as the breadth and depth grow, the scale of dissemination tends to increase exponentially. Therefore, a higher forwarding score indicates a wider and more extensive spread of the message. This observation provides further validation of the effectiveness of our method. Although the forwarding score was not explicitly factored into our estimation process, the estimated values of breadth and depth still successfully capture the characteristic that messages forwarded more frequently tend to propagate more widely and deeply. This consistency suggests that our method effectively reflects the structural dynamics of message dissemination, thereby enhancing its robustness and reliability in real-world scenarios.

## C Population-Level Estimates Based on Network Cascade

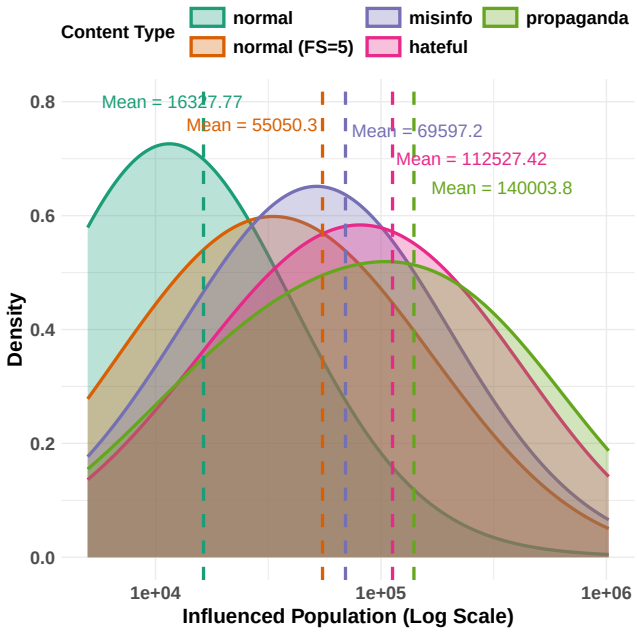



**Figure 6: Density of influenced population by content type based on network cascade. Vertical dashed lines indicate the mean influenced population for each content type.**

As shown in Figure 6, which provides population-level estimates based on the network cascade, the number of individuals affected by different types of messages is presented. The results are consistent with those obtained from the influence cascade analysis, further validating the robustness of our findings. Similarly to the influence cascade, harmful messages demonstrate a significantly larger reach compared to normal messages. Harmful messages, such as propaganda and hateful speech, continue to affect more people, while misinformation has a relatively smaller impact. This consistency between the two models strengthens the reliability of our results and highlights the structural differences in how harmful messages are disseminated among different populations.

# D Comparison of Structural Parameters for Influence and Network Cascades Across Different Demographic Features

**Table 7: Comparison of Structural Parameters for Influence and Network Cascades Across Different Demographic Features**

| Category | $\mu_b$ | $\sigma_b$ | $\mu_h$ | $\sigma_h$ | $\mu_k$ | $\sigma_k$ |
|---|---|---|---|---|---|---|
| **Income (Network Cascade)** | | | | | | |
| <25k | 2.72 | 0.253 | 4.53 | 0.340 | 1.00 | 0.048 |
| >25k | 2.71 | 0.195 | 4.48 | 0.233 | 1.00 | 0.042 |
| **Income (Influence Cascade)** | | | | | | |
| <25k | 2.79 | 0.382 | 4.47 | 0.164 | - | - |
| >25k | 2.73 | 0.292 | 4.44 | 0.1255 | - | - |
| **Religion (Network Cascade)** | | | | | | |
| Hinduism | 2.72 | 0.248 | 4.53 | 0.335 | 1.00 | 0.045 |
| Islam | 2.71 | 0.223 | 4.50 | 0.273 | 1.00 | 0.039 |
| **Religion (Influence Cascade)** | | | | | | |
| Hinduism | 2.78 | 0.380 | 4.47 | 0.163 | - | - |
| Islam | 2.75 | 0.330 | 4.46 | 0.141 | - | - |
| **Caste (Network Cascade)** | | | | | | |
| General | 2.74 | 0.254 | 4.54 | 0.344 | 1.00 | 0.054 |
| Other | 2.70 | 0.233 | 4.50 | 0.304 | 1.00 | 0.034 |
| **Caste (Influence Cascade)** | | | | | | |
| General | 2.80 | 0.390 | 4.48 | 0.167 | - | - |
| Other | 2.76 | 0.352 | 4.46 | 0.151 | - | - |
| **Education (Network Cascade)** | | | | | | |
| <high school | 2.71 | 0.234 | 4.51 | 0.315 | 1.00 | 0.038 |
| >high school | 2.72 | 0.253 | 4.53 | 0.331 | 1.00 | 0.049 |
| **Education (Influence Cascade)** | | | | | | |
| <high school | 2.77 | 0.366 | 4.46 | 0.157 | - | - |
| >high school | 2.79 | 0.371 | 4.47 | 0.160 | - | - |
| **Age (Network Cascade)** | | | | | | |
| >25 | 2.74 | 0.252 | 4.54 | 0.329 | 1.00 | 0.025 |
| 25-34 | 2.76 | 0.282 | 4.56 | 0.354 | 1.00 | 0.064 |
| >35 | 2.69 | 0.219 | 4.49 | 0.307 | 1.00 | 0.045 |
| **Age (Influence Cascade)** | | | | | | |
| >25 | 2.80 | 0.368 | 4.48 | 0.158 | - | - |
| 25-34 | 2.84 | 0.425 | 4.49 | 0.182 | - | - |
| >35 | 2.73 | 0.344 | 4.45 | 0.148 | - | - |