WhatsViral: A Privacy-Preserving System for Studying Viral Content on WhatsApp

Kiran Garimella
Rutgers University
kiran.garimella@rutgers.edu

Abstract

Studying the spread of content on end-to-end encrypted platforms like WhatsApp presents a significant challenge without compromising user privacy. This paper introduces a novel, privacy-preserving system to identify viral images and videos. Our approach utilizes an on-device application that locally scans a user's media gallery and computes hashes of content. Only these non-reversible hashes are transmitted to a server, where they are aggregated to detect widespread content. The full content is exported from the user's device only after it has been identified as viral by the system, ensuring user data remains private by default. We successfully deployed this application to hundreds of users in India, creating a unique dataset to analyze the network structure and characteristics of viral content spread on WhatsApp.

Our analysis reveals that "virality" is not a monolithic phenomenon; we deconstruct it into two empirically decoupled dimensions: breadth (widespread prevalence across the network) and depth (long forwarding chains). We provide the first large-scale, empirical evidence that the most common form of virality—high breadth—is driven by complex contagion, requiring social reinforcement within dense communities to spread. Computational simulations on a reconstructed network validate this framework, demonstrating that only a complex contagion model can reproduce the observed widespread prevalence.

Our work makes a dual contribution: we provide a new, ethical paradigm for studying encrypted ecosystems and use it to reveal a more nuanced model of information diffusion, exposing significant blind spots in current content moderation policies that focus exclusively on forwarding depth.

1 Introduction

End-to-end encrypted platforms like WhatsApp are widely used by billions of people across the world. These encrypted ecosystems, however, are built on a core tension: the very encryption that empowers users with vital privacy also renders the platform's content opaque to any form of external moderation. This creates a critical blind spot for society. While journalists and researchers have documented the severe real-world harms linked to viral content on WhatsApp, from the coordination of lynchings in India [1] to the rampant spread of health misinformation during global pandemics [8, 16]. We lack the basic methodological tools to systematically study what becomes popular, how it spreads, and why. Understanding these dynamics is of interest beyond just an academic audience, and a solution to this could help maintaining information integrity and public safety on WhatsApp.

The difficulty of this problem is rooted in the technical and ethical guarantees of end-to-end encryption. Unlike public social networks where content can be observed, on WhatsApp, not even the platform itself can access message content. This makes traditional, server-side content analysis and moderation impossible by design. Any naive approach that proposes a "backdoor" [23] or server-side content scanning would fundamentally compromise the platform's private design, destroying user trust and creating security vulnerabilities [26]. The challenge, therefore, is to find a way to gain macroscopic insights into population-level content trends without compromising the microscopic privacy of individual conversations. This requires a paradigm shift away from centralized data collection and towards methods that respect the principles of data minimization and user control.

Prior research has attempted to navigate this challenging landscape with innovative but ultimately limited methodologies. One common approach involves scraping data from publicly accessible WhatsApp groups [12, 25]. While useful, this method captures the behavior of a small, non-representative subset of users and fails to shed light on the private, high-trust peer-to-peer and small-group chats that constitute the vast majority of activity on the platform. A second approach involves data donation tools, where consenting participants export their chat archives for analysis [10]. These studies provide incredibly rich, deep datasets but face an "all-or-nothing" privacy trade-off where users consent to share *every* piece of content that was shared in certain WhatsApp group, which creates a high barrier to participation. This limits the scale and diversity of samples and makes such methods better suited for deep qualitative analysis than for measuring the broad, population-level prevalence of any given piece of content. A scalable, low-friction method for measuring content prevalence without requiring users to surrender their entire private archives has remained an unsolved problem.

This paper introduces a novel system that directly addresses this methodological gap. We designed, built, and deployed a privacy-preserving, on-device application that identifies viral images and videos on WhatsApp. Our approach is a form of edge computing: the application resides on a user's phone and performs all sensitive operations locally. It periodically scans the device's WhatsApp media folder, computes non-reversible perceptual hashes of the content on-device, and transmits only these anonymous, compact fingerprints to a central server. By aggregating these hashes from hundreds of participants, our system can detect when a piece of content has achieved widespread prevalence. Only after a content hash has been flagged as "viral" by the server is a request sent back to the participating devices to export the full media file for analysis. This privacy-by-design architecture ensures that personal, non-viral media never leaves a user's device and is never seen by the research team.

Using this system, we provide the first large-scale, prevalence-based view of viral content on WhatsApp. Our analysis yields three key contributions. First, methodologically, we present a scalable

and ethical paradigm for studying encrypted platforms. Second, theoretically, we deconstruct the monolithic concept of "virality," demonstrating empirically that widespread prevalence (breadth) is a distinct phenomenon from long forwarding chains (depth), with each mode of spread characterized by different types of content. Finally, we provide powerful, direct evidence that the most common form of virality—widespread community saturation—is governed by the dynamics of complex contagion, requiring social reinforcement from multiple peers. This work not only provides a new tool for safely illuminating "dark social" platforms but also reveals a more nuanced, socially-grounded understanding of how information truly diffuses through the world's largest communication network.

2 Related work

Our research contributes to a conversation spanning four distinct but interconnected domains: (1) foundational theories of social contagion; (2) empirical studies of virality on public platforms; (3) the methodological challenges of studying encrypted "dark social" spaces; and (4) the development of privacy-enhancing technologies for data analysis.

Foundational Theories of Social Contagion. The study of diffusion is rooted in a theoretical distinction between simple and complex contagions. Simple contagions model the spread of information that requires minimal social proof for adoption, such as a novel fact or a piece of breaking news. Foundational network theory suggests that this type of spread is best facilitated by "weak ties"—long-range social bridges that connect otherwise distant clusters within a network [14]. Mathematical models like the Independent Cascade (IC) model formalize this process, where an activated node has an independent, probabilistic chance to activate its neighbors, allowing information to traverse a network via single exposures [17].

In contrast, complex contagions model the adoption of behaviors or norms that are perceived as risky, costly, or illegitimate, and thus require social reinforcement from multiple sources. Sociological theories posit that for these contagions, exposure from a single contact is insufficient; adoption requires a threshold of confirmation from one's local neighborhood [4]. This process is mathematically captured by models like the Linear Threshold (LT) model, where a node activates only after a sufficient fraction of its neighbors have become active [17]. Consequently, complex contagions are not facilitated by sparse bridges but rather by dense, clustered network structures with high triadic closure, which provide the necessary channels for social reinforcement. Empirical work has confirmed these dynamics in contexts ranging from the adoption of health behaviors to participation in online forums [3]. Further research has refined this by introducing the concept of "structural diversity," finding that exposure from multiple, disconnected social contexts can be a more powerful trigger for adoption than repeated exposure from a single, homogenous cluster [30].

Empirical Studies of Virality on Public Platforms. The advent of public social media platforms like Twitter, Facebook, and YouTube enabled the first wave of large-scale, data-driven research on virality. This work moved beyond theory to empirically map the structures and features of viral events. Structurally, studies revealed that massive cascades are rarely the product of long, unbroken chains. Instead, they are typically broad and shallow, driven by a critical mass of initial broadcasts that trigger many smaller, independent cascades [13]. This led to a re-evaluation of the "influential hypothesis"; while influencers can be important seeds, large-scale spread often depends more on a population of easily influenced "ordinary" users who propagate content within their communities than on a few elite individuals [2, 31].

Research into viral content has identified several key characteristics. Emotion is a powerful driver, particularly high-arousal emotions like awe, anger, and anxiety, which stimulate the physiological urge to share [21]. Content that is surprising, practically useful, or contains strong narrative elements also shows higher potential for diffusion. However, despite these insights, the literature has consistently affirmed the profound difficulty of predicting virality. The process is highly stochastic, with nearly identical content often leading to vastly different outcomes, suggesting that timing, luck, and network context play an enormous role [5, 27].

Our work contributes to this empirical literature not by attempting to predict virality, but by providing a more nuanced framework for explaining its outcomes. We extend the structural finding that virality is broad and shallow by creating a formal breadth vs. depth typology. This allows us to show that "broadcast virality" and "niche virality" are distinct phenomena with different content signatures (e.g., local/utilitarian vs. national/ideological). By linking these patterns to their underlying contagion mechanisms, we offer a deeper explanatory model for why different types of content spread in different ways, moving beyond correlation-based feature analysis.

The Methodological Challenge of "Dark Social". While public platforms offer rich data, a significant and growing portion of digital communication occurs in "dark social" spaces—fully encrypted, private messaging applications like WhatsApp, Signal, and Telegram. These platforms present an acute methodological dilemma: they are crucial sites for the spread of information and misinformation, yet their end-to-end encryption makes them largely inaccessible to traditional research methods [28]. The societal importance of overcoming this hurdle is immense, as misinformation campaigns on these platforms have been tied to severe real-world harms, including political violence and public health crises [1, 6].

Researchers have developed several strategies to navigate this challenge, each with its own trade-offs. One approach is to study the content of public or semi-public groups on these platforms, but this risks capturing the behavior of a highly engaged and unrepresentative minority of users [12, 25]. Another recent method involves data donation, where consenting users contribute their chat archives for analysis [10, 11]. This provides unparalleled, high-fidelity data but is logistically complex, raises significant privacy and data management challenges, and is often limited to a smaller number of highly motivated participants. It is optimized for deep, qualitative analysis rather than a broad, population-level measure of content prevalence.

Our work introduces a new methodological paradigm that directly addresses the limitations of prior approaches. We are not studying a non-representative public subset, nor are we requiring a full donation of private archives. Our on-device, hash-based system is a "light-touch" alternative designed specifically to measure

content prevalence at scale, without collecting any non-viral or personal content by default. It provides a complementary, macroscopic view of the information ecosystem that was previously unattainable, demonstrating a path to measure broad diffusion patterns while respecting the principles of end-to-end encryption.

Privacy-Enhancing Technologies for Data Analysis. Our technical approach is informed by the principles of Privacy-Enhancing Technologies (PETs), a field dedicated to deriving insights from data while minimizing privacy risks. A prominent example is Federated Learning, which enables the training of machine learning models on decentralized data stored on user devices, avoiding the need to collect raw data centrally [20]. Another formal framework is Differential Privacy, which provides a mathematical guarantee that the output of an analysis will not meaningfully change if any single individual's data is removed, often achieved by adding calibrated statistical noise to results [7]. Other approaches rely on cryptography, such as Secure Multi-Party Computation [18], which allows multiple parties to jointly compute a function over their inputs without revealing those inputs to each other.

However, these techniques are primarily designed for training predictive models or answering pre-defined aggregate queries, making them less suited for open-ended, discovery-oriented tasks like identifying previously unknown viral content. Furthermore, the computational overhead of many cryptographic methods is impractical for continuous monitoring on low-end mobile devices. Our system addresses this gap by representing a novel, practical application of privacy-preserving principles to the discovery problem. We use a lightweight, on-device hashing mechanism that allows a server to learn that content is popular only after it has reached a prevalence threshold, a form of results-based privacy.

3 WhatsViral App Design

Our primary contribution is the design and deployment of WhatsViral, an Android application that enables the study of media diffusion while rigorously protecting user privacy. The system's architecture is predicated on a key feature of WhatsApp: media files are stored directly on a user's device in a designated folder. Our application leverages this on-device storage to analyze content prevalence without ever accessing private communications or breaking encryption.

The entire system was engineered with two guiding principles: privacy by default and resilience on low-end hardware common in regions like India. The application was developed natively for Android in Kotlin, using modern Android Jetpack libraries to manage robust background tasks (`WorkManager`) and local data persistence (`Room`). The data collection workflow, illustrated in Figure 1, proceeds in several automated steps:

- On-Device Media Scanning: A background service on the user's device periodically scans the dedicated WhatsApp media folders for new or previously unseen images and videos.
- Local, Privacy-Preserving Hashing: For each media file, the application computes a compact, non-reversible fingerprint directly on the device. Crucially, the raw content never leaves the user's phone at this stage. We use the robust PDQ perceptual hash for images, which allows us to group visually similar variants of the same content (e.g., with different watermarks or compression levels) [9]. For videos, we use a standard MD5 hash. This choice was a deliberate engineering trade-off; more sophisticated video perceptual hashes were computationally too expensive to run reliably on the low-end smartphones prevalent in our study population.
- Resilient Synchronization: The generated hashes are first stored in a local on-device database. This ensures that data collection continues even if the user is offline. When a network connection is available, the application uploads only the queue of anonymized hashes to our central server.
- Server-Side Aggregation and Viral Detection: The server receives a stream of anonymous hashes from all participants and aggregates them. A content item is flagged as 'viral' once its hash appears in the uploads from a predefined number of unique devices (k=5 in our case).
- Triggered Content Export: Only after a hash is confirmed to be viral by the server does the system complete the privacy-preserving loop. The server sends a request back to the specific devices that hold the viral content, prompting the application to then export the full media file for our thematic analysis. This ensures that personal, non-viral media is never seen by the research team.

This multi-stage process represents a significant engineering effort to balance the demands of rigorous data collection with the ethical imperative of user privacy, creating a viable methodology for safely studying encrypted ecosystems.

4 Data collection

This study is based on three distinct types of data collected from a single cohort of participants in Uttar Pradesh, India: (1) media prevalence data from our custom WhatsViral application, (2) on-device contact networks, and (3) ground-truth forwarding data from the WhatsApp Explorer tool [10].

Participant Recruitment. We recruited 547 participants through an in-person, network-based sampling approach in villages surrounding Lucknow and Shahjahanpur. The process was initiated by 15 local surveyors who first enrolled their direct contacts. Participants were then encouraged to invite their friends, creating a snowball sample that reflects real-world social structures. All participants provided informed consent and were compensated 600 Indian rupees (approx. $7 USD) for keeping the WhatsViral app installed for a minimum of seven days. The resulting participant pool was predominantly male (88.2%) with a mean age of 28.5 years.

The WhatsViral Dataset. The primary dataset was collected via our privacy-preserving WhatsViral application. The app scanned each participant's on-device WhatsApp media folder and uploaded non-reversible hashes of their images and videos. In total, we collected 294,400 image and 14,024 video hashes. From this corpus, we identified 10,417 unique items (9,213 images, 1,204 videos) as 'viral'—defined as content present on the devices of at least five participants. A key feature of this on-device approach is its ability to retrospectively identify content; the oldest viral item detected in our dataset dated back to August 13, 2022, long before our study began.

The Contact Network Dataset. To map the social fabric of our participants, the WhatsViral app also collected their contact lists

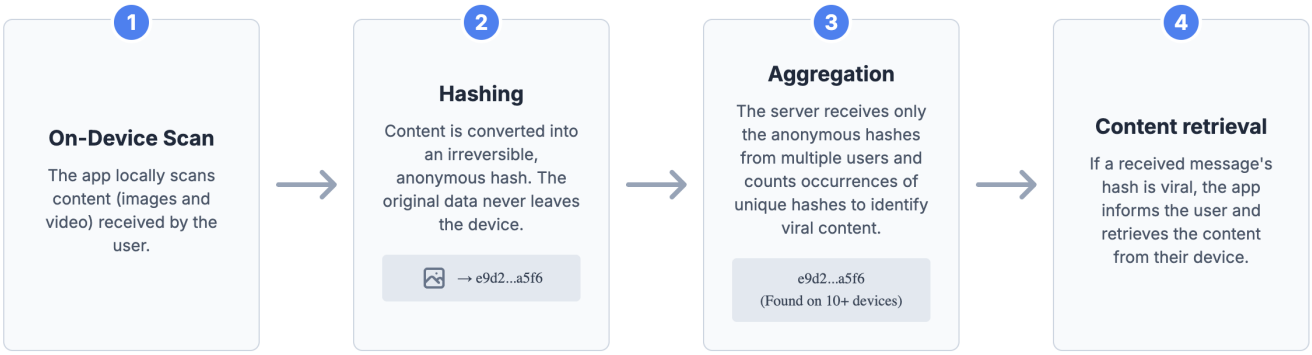

Figure 1: The end-to-end data collection and analysis workflow of the WhatsViral system. The process is designed to be privacy-preserving by default, with on-device hashing and server-side aggregation. Full content is only exported from a device after its corresponding hash has been identified as viral by the central server.

(phone numbers on the contacts app of a participant). To protect privacy, all phone numbers were hashed on-device before being transmitted, meaning no raw contact information ever left the user's phone. Due to technical issues on certain devices (see Section 6), we successfully collected contact networks from 461 of the 547 participants. These users maintained extensive social networks, with an average of 415 contacts each (median = 234, max = 3,455). This raw data served as the foundation for constructing the ego-overlap network used in our analysis.

From this data, we constructed an ego-overlap network, where each participant is a node and an edge between two nodes signifies a strong social tie, weighted by the number of mutual contacts. The resulting network of 461 nodes and 26,507 edges was highly connected, with its largest connected component encompassing 98% of all participants. This network was also characterized by its high degree of local structure, or "cliquishness." We measured a global clustering coefficient of 0.833, indicating a strong tendency for triadic closure (i.e., the friends of a user are also likely to be friends with each other). To put this in perspective, an Erdős-Rényi (ER) random graph of similar size and density has a clustering coefficient of only 0.435. This significant difference confirms that our sample contains a truly social network within a densely clustered social fabric. This property is a key prerequisite for the social reinforcement mechanisms we investigate later in this paper.

The WhatsApp Explorer Dataset (Ground-Truth Forwarding). To provide a ground-truth baseline for virality based on forwarding, we recruited a random subsample of 54 users (10% of our cohort) to donate their data using the WhatsApp Explorer tool [10]. These users consented to export the contents of their WhatsApp groups, donating an average of 4.5 groups each for a total of 245 groups providing over 67,000 messages . This dataset provided access to message metadata, including WhatsApp's 'forwarded many times' flag, which is applied to content that has been forwarded through a chain of at least five hops from the original sender [32]. Roughly 1% of the WhatsApp Explorer messages (689 messages) were marked as 'forwarded many times'. Such forwarded many times content is usually considered by previous studies [11] and WhatsApp as 'viral', and hence we consider this set as a rough ground truth for what was spreading virally on WhatsApp.

5 Analysis

Based on these rich, complementary datasets, our analysis proceeds in five stages. We beging by conducting a thematic analysis of the viral content identified by WhatsViral, compaing it with the viral content labeled by WhatsApp Explorer, revealing key differences between prevalence-based and forwarding-based virality. This leads us to deconstruct the concept of virality into two distinct dimensions: breadth and depth. We then investigate the underlying mechanism of spread, providing strong empirical evidence for complex contagion. Finally, we validate our entire framework through computational simulations on a large-scale reconstructed network, demonstrating that social reinforcement is necessary to produce the widespread content prevalence we observe in the wild.

5.1 Thematic Analysis of Viral Content

Our initial analysis sought to characterize the content identified as 'viral' by our prevalence-based WhatsViral system. To provide a robust baseline, we compared this content against a ground-truth dataset of media marked by WhatsApp as 'forwarded many times'—a standard proxy for virality in previous studies [11].

First, we identified a substantial cohort of media present in both datasets, matching 3,809 image clusters and 153 video clusters[1] by their content hashes. We then employed a large multimodal AI model (Gemini 2.5 Pro) to systematically analyze these items from both sources. For each piece of content, the model classified its topics, identified any prominent public figures, and generated a detailed description. This comparative analysis revealed stark thematic differences between the two datasets, suggesting that prevalence-based virality and high-forwarding virality capture distinct types of popular content.

The most significant divergence appears in the dominant topics, as illustrated in Figure 2. Content achieving high prevalence in the

[1]A 'cluster' comprises visually similar variants of the same media, accounting for modifications like cropping, watermarks, or compression.

WhatsViral dataset is overwhelmingly local and utilitarian. The most frequent categories include 'Community & Local Life' (17.7%), 'Culture & Heritage' (17.9%), and 'Education & Youth' (17.1%). In contrast, content from the WhatsApp Explorer dataset, while also featuring community topics, skews significantly more toward national 'Politics & Governance' (14.7%) and 'Religion & Spirituality' (17.2%). This distinction is further highlighted by niche topics; for instance, agricultural content was more than twice as common in the WhatsApp Explorer set (8.2% vs. 3.7%).

This local-versus-national pattern is also reinforced by the public figures identified within the images. The WhatsApp Explorer content is dominated by top-tier national politicians, with Prime Minister Narendra Modi appearing in 4.8% of items, followed by Yogi Adityanath (3.6%) and Amit Shah (1.2%). While these national figures also appear in the WhatsViral dataset, they are featured less frequently and share the stage with a variety of local leaders (e.g., Satish Chandra Sharma, 1.8%).

Taken together, these findings paint a clear picture. The content that becomes widely prevalent (WhatsViral) is often practical, community-oriented, and locally relevant. The content that travels along long forwarding chains (WhatsApp Explorer) is more frequently ideological, political, and national in scope. This divergence suggests that we are observing two fundamentally different modes of information diffusion, a concept we will deconstruct in the next section.

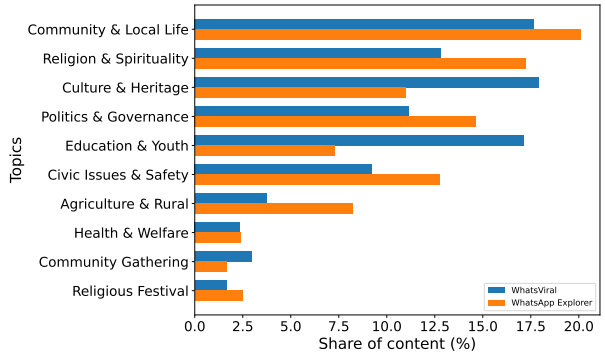

Figure 2: Comparison of dominant topics in content from the WhatsViral and WhatsApp Explorer datasets. Bars represent the percentage of items in each dataset classified under a given topic.

5.2 Deconstructing Virality: Breadth vs. Depth

The thematic differences detailed in the previous section suggest that the popular content identified by our prevalence-based system (WhatsViral) is fundamentally different from the content identified by forwarding chains (WhatsApp Explorer). This leads to a critical question: are these two measures of "virality"—widespread adoption versus long forwarding paths—simply different proxies for the same underlying phenomenon, or do they represent distinct modes of diffusion?

To answer this, we formally tested the relationship between these two dimensions for the 3,962 content clusters present in both datasets. We classified each item along two axes: (i) Breadth: Whether an item was viral by prevalence in the WhatsViral dataset. (ii) Depth: Whether an item had a high forwarding score (forwarded in a chain of at least 5 hops from the original sender) in the WhatsApp Explorer dataset, indicating a long transmission path.

The analysis reveals a striking decoupling between these two metrics. As shown in Table 1, the vast majority of content that achieves high breadth does not travel with great depth, and vice versa. Of the 535 items with high breadth, 98.5% (527) had low depth. Similarly, of the 87 items with high depth, 90.8% (79) failed to achieve high breadth. This near-orthogonality confirms that breadth and depth are not interchangeable measures of virality; they are distinct phenomena.

Table 1: Virality Matrix: Number of content clusters classified by Breadth (Prevalence) and Depth (Forwarding Path).

	Low Depth	High Depth	Total
Low Breadth	3,348	79	3,427
High Breadth	527	8	535
Total	3,875	87	3,962

This empirical finding motivates a more nuanced framework that treats virality as a two-dimensional construct. Breadth measures the total number of people who possess an item, capturing its final reach or saturation. Depth measures the topological length of a single transmission chain, capturing its sequential persistence. Content can achieve one without the other; for example, a "broadcast" from an influential source results in high breadth but low depth, while a "niche rumor" passed along a long chain of weak ties results in high depth but low breadth. This framework yields a four-part typology of content diffusion. To characterize the content within each category, we applied the thematic analysis from Section 5.1, using the topics generated by the Gemini model. This reveals a distinct topical signature for each of the four diffusion patterns:

- High Breadth, Low Depth (Broadcast Virality): This is the most common form of viral content in our study (527 items). It consists of the locally-relevant, utilitarian content identified in the previous section, such as educational materials, local news, and government circulars that spread widely within a community but are not sequentially forwarded many times.
- Low Breadth, High Depth (Niche Virality): These 79 items travel far but not wide. Thematically, this content is often identity-based, such as religious greetings or cultural memes that are passed along specific social corridors without breaking into the wider population.
- High Breadth, High Depth (Epidemic Virality): A rare but powerful category (8 items) that excels on both dimensions. This content is typically characterized by high urgency and broad appeal, such as videos of accidents or natural disasters.
- Low Breadth, Low Depth (Unsuccessful Content): The vast majority of content (3,348 items) fails to propagate significantly along either axis. Since this corresponds to almost 85% of the content, the categories of content here reflect the overall categories described in Section 5.1.

This framework moves the research question from if content is viral to in what way it is viral. It provides the theoretical lens through which we can now analyze the underlying mechanisms driving these different diffusion patterns.

5.3 The Mechanism of Spread: Social Ties and Reinforcement

Having established that virality on WhatsApp manifests in distinct "broadcast" (high-breadth) and "niche" (high-depth) patterns, we now investigate the underlying mechanism driving this diffusion. A central question in information spread is whether content propagates primarily through the structure of personal relationships or through broadcast dynamics in large, impersonal settings (e.g., groups with many strangers). Groups on WhatsApp, especially large public groups with strangers are widely used with over one in five users reporting to be a part of at least one such group in Brazil and India [19, 22]. We hypothesized that content diffusion on WhatsApp is fundamentally a social process, where the likelihood of two individuals being exposed to the same content is a direct function of their social proximity.

To test this, we leveraged the ego-overlap network introduced previously, where the strength of a social tie is measured by the number of mutual contacts two users share. We segmented all pairs of users in our network into three distinct categories: (i) Strong Ties: Pairs with an above-median number of mutual contacts. (ii) Weak Ties: Pairs with a below-median (but non-zero) number of mutual contacts. (iii) Strangers: Randomly sampled pairs of users with no connecting edge in our network. For each category, we calculated the mean content overlap: the probability that two users in a pair had both received the same unique content item.

The results, shown in Figure 3, provide decisive support for our hypothesis. Pairs of users connected by strong ties exhibited the highest mean content overlap. This value, while small in absolute terms, is an order of magnitude greater than the overlap for users connected by weak ties. Most tellingly, the content overlap between strangers was negligible, approaching zero.

This steep gradient provides compelling evidence that information spread on WhatsApp is not a process of random broadcast. The probability that two people see the same content is strongly conditioned by the existence and strength of their social connection. The near-zero overlap between strangers argues against the theory that large, anonymous groups are the primary amplifiers of content. Instead, the data points to a model of social contagion where content flows along the established pathways of friendship.

This finding, however, raises a more specific question: is a single exposure from a friend sufficient for content to spread (a simple contagion), or is reinforcement from multiple social contacts required (a complex contagion)? We address this critical distinction in the following analysis.

5.4 Empirical Evidence for Complex Contagion

To empirically test the hypothesis of complex contagion, we conducted a "dose-response" analysis designed to measure the effect of social reinforcement on content adoption. The core thesis of complex contagion is that the probability of adoption is a non-linear,

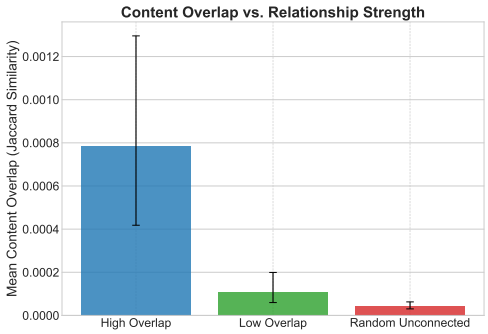

Figure 3: Mean content overlap between pairs of users as a function of tie strength. Users with strong & weak social ties have significantly higher content overlap than randomly selected pairs of strangers, whose overlap is near-zero.

accelerating function of the number of peers who have already adopted—a pattern that requires multiple reinforcing signals.

For constructing the dose-response curve we created a time-ordered event log for each unique content item by merging adoption timestamps with the network data. An exposure event occurs for a user u at time t when one of their neighbors adopts a piece of content, incrementing their total number of exposed neighbors, or "dose," to k. For every such exposure event, we record an outcome for user u. If u adopts the same content at or shortly after time t, we log an "adoption event" (a success) at dose k. If u does not adopt the content upon that exposure, we log a "non-adoption event" (a failure) at dose k. By aggregating the counts of all success and failure events across all users and content items, we calculated the empirical conditional probability of adoption for each dose k.

The resulting empirical dose-response curve, plotted in Figure 4, provides a clear visual signature of complex contagion. The probability of adoption is vanishingly small for a single exposure ($P(\text{adopt}|k{=}1) \approx 3.1 \times 10^{-4}$). The curve begins to climb sharply at k=3, where the probability increases by an order of magnitude. This distinct sigmoidal ("S-shaped") profile, characterized by an initial period of resistance followed by accelerating adoption, is inconsistent with a simple contagion model and strongly indicative of a social reinforcement mechanism.

To formalize this visual evidence, we fit a logistic regression model to predict the log-odds of adoption based on the exposure dose k. To test for the hypothesized non-linearity, we included both a linear and a quadratic term for k. The model is specified as: $\text{logit}(P(\text{adopt})) = \beta_0 + \beta_1 k + \beta_2 k^2$.

The results of the regression analysis were unambiguous. We found a highly significant positive coefficient for the linear term (β_1=0.78, $p < 10^{-15}$) and a highly significant negative coefficient for the quadratic term (β_2=-0.0318, $p < 10^{-15}$). The positive linear term confirms that more exposures increase adoption odds, while the significant quadratic term provides robust statistical evidence that this relationship is non-linear, capturing the accelerating-then-saturating effect of reinforcement that defines complex contagion.

These findings provide clear evidence that content diffusion on WhatsApp is governed by the dynamics of a complex contagion.

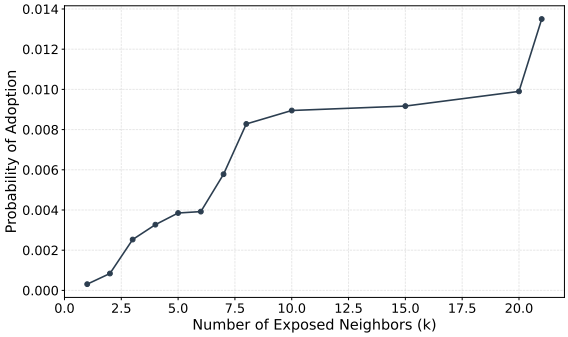

Figure 4: The probability of content adoption as a function of the number of infected neighbors (k). The curve exhibits a clear sigmoidal shape, where the probability of adoption is negligible for a single exposure but accelerates as the number of reinforcing social signals increases, providing strong evidence for complex contagion.

Adoption is not a simple function of exposure; it requires a threshold of social proof from multiple peers. This mechanism is the key to understanding the high-prevalence "broadcast" virality we identified earlier. Content becomes widespread not by being randomly broadcast to strangers, but by successfully achieving critical mass within the dense, clustered communities of the social network, triggering cascades of reinforcement-based adoption.

5.5 Validating Contagion Dynamics via Simulation

Our empirical findings distinguish between path-based (depth) and prevalence-based (breadth) virality, with evidence suggesting the latter is driven by complex contagion. To validate this theoretical framework, our final analytical step uses computational simulation. The goal is to test whether a complex contagion mechanism is necessary to reproduce the widespread prevalence observed in our empirical data. While our observed ego network captures a realistic local WhatsApp network, its limited size is insufficient for studying large-scale diffusion. Therefore, to test our hypotheses in a generalizable, large-world scenario, we first generated a large-scale synthetic network that faithfully reflects the properties of our participant population.

5.5.1 Inferring a Realistic, Large-Scale Social Network. To generate a realistic network from our egocentrically sampled data, our process was twofold:

First, we generated a structurally plausible "skeleton" network. After correcting for friendship-paradox bias in our data to produce an accurate degree distribution, we trained a dyad-independent Exponential Random Graph Model (ERGM) surrogate. This involved featurizing each pair of users with metrics like mutual contact overlap, degree parity, and homophily, and fitting a stabilized logistic model that learned the rules of tie formation with approximately 92% accuracy. This model was then used to generate a 10,000-node graph that faithfully preserved these learned structural properties.

Second, we augmented this skeleton network with content-based weak ties to capture the long-range bridges crucial for diffusion. We scored potential inter-community links using a metric that combined the Jaccard similarity of users' shared content with a weighting for content rarity (Jaccard×(1+rarity)). By injecting the top 5% of these predicted ties into the network, we dramatically increased its global clustering and cohesion.

The final hybrid network of 10,000 nodes and 54,000 edges provided a robust and realistic substrate for simulation, exhibiting both the dense local communities and the inter-community bridges characteristic of real-world social networks.

5.5.2 Simple vs. Complex Contagion Dynamics. On this reconstructed network, we simulated the two distinct contagion models that correspond to our virality framework. The outcomes were dramatically different: (i) Simple Contagion (Path-Based Virality): To model high-depth, niche content, we used an Independent Cascade (IC) model [17] with a transmission probability of p=0.12. Across 500 simulations, this simple contagion process activated an average of 2,545 nodes (26% of the network's giant component). The cascades saturated quickly, consistently failing to break out of initial community clusters. (ii) Complex Contagion (Prevalence-Based Virality): To model high-breadth, widespread content, we used a Linear Threshold (LT) model (θ ~U[0,1]) [17], which operationalizes social reinforcement. In stark contrast, this complex contagion process activated an average of 8,882 nodes (91% of the giant component). Once a cascade reached a critical mass within a few clustered neighborhoods, the reinforcement effect triggered explosive, network-spanning adoption.

The nearly four-fold difference in final cascade size is the key result. It demonstrates that the network's structure disproportionately benefits a reinforcement-based diffusion process. This provides a clear, mechanistic explanation for the widespread "broadcast" virality seen in our empirical data. These simulation results are robust to variations in model parameters (e.g., p, θ), MCMC simulation steps, network size, and initial seed set sizes; the vast gap in reach between the IC and LT models is a structural feature of the system, not an artifact of our choices. Table 2 (Appendix) show the robustness results for an exhaustive search of parameter space across five different variables.

This final analysis closes the loop on our analysis. It demonstrates that the two distinct forms of virality observed in our data are explained by different diffusion mechanisms operating on a realistic social fabric. Path-based (depth) virality is consistent with a simple contagion process that spreads through chains but is ultimately contained. In contrast, the widespread prevalence (breadth) that characterizes the most visible content on WhatsApp is reproducible only through a complex contagion model, where social reinforcement within clustered communities drives massive, cascading adoption. This validates our central thesis: to understand information spread in encrypted ecosystems, one must deconstruct the monolithic concept of "virality" and account for the powerful, underlying dynamics of complex contagion.

6 Discussion

To study content propagation and virality within an encrypted ecosystem, we developed and deployed a novel, privacy-preserving tool for data collection on WhatsApp—a platform that is used by over 2 Billion users but has largely remained opaque to researchers.

This approach yielded a unique dataset of shared media, enabling us to move beyond conventional metrics and identify the core mechanics of virality. Our analysis reveals a fundamental distinction between two modes of viral spread: widespread prevalence (breadth), which we find is driven by social reinforcement within dense communities (complex contagion), and long forwarding chains (depth). Crucially, we discovered that the most common form of virality—the broad, community-level saturation of content—is governed by a mechanism that is currently invisible to platform moderation policies focused on forwarding limits. Our findings expose a critical blind spot in current content moderation strategies on platforms like WhatsApp. Policies designed to limit the forwarding of messages [32] are exclusively targeting depth-based virality. While this may curb the spread of niche rumors, it does nothing to address the far more common phenomenon of broadcast virality.

As our simulations confirm, content driven by complex contagion can reach nearly the entire network without a single message being "forwarded many times" [15]. A malicious actor, such as a political party seeking to spread propaganda [24, 25, 29], does not need long forwarding chains. Instead, they can employ a large number of agents to "seed" content within many social clusters. The natural dynamics of social reinforcement will then ensure the content achieves massive breadth, completely bypassing current moderation tripwires. Future governance policies must account for the mechanics of broadcast and complex contagion to be effective.

Perhaps the most significant contribution of this work is methodological. Studying encrypted platforms without compromising user privacy is one of the greatest challenges in computational social science. Our on-device, hash-based system presents a viable edge-computing paradigm for ethical data collection. By performing analysis locally and only transmitting anonymized, non-reversible hashes to a central server, we can observe macroscopic diffusion patterns while user content remains private by default. As researchers and regulators grapple with the opacity of encrypted systems, the development of such privacy-preserving tools is not just a technical contribution, but a necessary step towards responsible oversight.

This methodological paradigm also inspires practical, privacy-preserving interventions platforms could adopt to mitigate widespread virality without breaking encryption. For instance, an application can introduce gentle friction on-device, such as a brief cooldown, when it detects a user has received the same media from many distinct contacts. It could also slow down broadcast bursts where a user sends the same content to numerous chats in a short period. Finally, platforms could implement a dual-label system, adding a breadth label like "Trending in your network"—computed with privacy-preserving techniques—to complement the existing depth label for forwarded content.

Limitations. Our study, while novel, has several limitations that provide avenues for future research. Our approach focuses on visual media, excluding text and links. Although prior work suggests images and videos constitute a large share of viral content on WhatsApp [12], they are not the whole story. Extending our privacy-preserving paradigm to text will require new on-device techniques, such as computing compact fingerprints from content embeddings before export. Furthermore, our prevalence counts are inherently conservative due to user behaviors like disabling media auto-download or deleting content to save storage. These factors likely result in a greater undercounting of larger video files than images. Future work could attempt to model this data sparsity or use device cache metadata to better bound its impact.

Our recruitment method also affects the generalizability of our findings. We used a network-based convenience sample, as our early attempts at broad recruitment through online ads and panels were costly and yielded little usable data. This experience, however, revealed a crucial lesson: to effectively study social contagion especially with a design like ours, on a platform like WhatsApp, the recruitment strategy must itself be social. We found that this kind of tool benefits most from seeded recruitment within real communities, facilitated by local partners who can provide sustained participant support. While this approach limits statistical generalizability, it is far more effective at capturing the tie-level network dynamics central to our analysis.

Practical Challenges. Deploying a mobile data collection tool in the real world presents formidable logistical and technical hurdles that are seldom reported in academic literature. Engineering constraints, particularly on the Android platform, significantly shaped the data we could collect. A primary design consideration was ensuring our application remained lightweight and functional on the low-end smartphones common in our study region. This reality imposed significant constraints on our methodology. For instance, while sophisticated perceptual hashes (e.g., Facebook's TMK video hashing [9]) are highly effective for video clustering, their computational cost made them infeasible for on-device execution without quickly draining battery or crashing the app on older hardware. Similarly, other potential on-device analyses, such as generating embeddings from text messages, had to be discarded as they would exhaust the limited memory and processing power of typical devices. Consequently, researchers must often trade methodological sophistication for practical feasibility, opting for simpler, less resource-intensive algorithms to ensure broad compatibility.

Another technical challenge stems from Android's fragmented ecosystem, where background work is notoriously fragile. Despite modern schedulers, many device manufacturers aggressively throttle or terminate background processes to conserve battery life—a practice that became a major source of data loss in our study. This issue was particularly acute for devices from manufacturers like Vivo, Oppo, Xiaomi, and Realme, which are popular in our study's region of India. Consequently, core data collection tasks, such as media hashing and contact network processing, failed for approximately 15% of our participants. Finally, the app store review process is a significant procedural hurdle. Applications requiring the expansive permissions necessary for background data collection face intense scrutiny from platforms like the Google Play Store. This can lead to lengthy review cycles and the risk of arbitrary, automated account suspensions, posing a significant threat to a project's timeline and viability.

Conclusion. This paper demonstrates that the spread of information on encrypted platforms is more complex and socially embedded than previously understood. By deconstructing "virality" into the distinct phenomena of breadth and depth, we reveal that the most prevalent content spreads through social reinforcement, a mechanism currently invisible to platform policies. Our work provides

both a new theoretical lens for understanding these crucial ecosystems and a novel, privacy-preserving methodology to study them ethically and effectively.

References

[1] Chinmayi Arun. 2019. On WhatsApp, rumours, lynchings, and the Indian Government. *Economic & Political Weekly* 54, 6 (2019).

[2] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. 2012. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*. 519–528.

[3] Damon Centola. 2010. The spread of behavior in an online social network experiment. *science* 329, 5996 (2010), 1194–1197.

[4] Damon Centola and Michael Macy. 2007. Complex contagions and the weakness of long ties. *American journal of Sociology* 113, 3 (2007), 702–734.

[5] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted?. In *Proceedings of the 23rd international conference on World wide web*. 925–936.

[6] Philipe de Freitas Melo, Mohamad Hoseini, Savvas Zannettou, and Fabrício Benevenuto. 2024. Don't break the chain: Measuring message forwarding on whatsapp. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 1054–1067.

[7] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science* 9, 3–4 (2014), 211–407.

[8] Azza El-Masri, Martin J Riedl, and Samuel Woolley. 2022. Audio misinformation on WhatsApp: A case study from Lebanon. *Harvard Kennedy School Misinformation Review* 3, 4 (2022), 1–13.

[9] Facebook. 2018. GitHub - facebook/ThreatExchange: Trust & Safety tools for working together to fight digital harms. — github.com. https://github.com/facebook/ThreatExchange. [Accessed 07-10-2025].

[10] Kiran Garimella and Simon Chauchard. 2024. WhatsApp explorer: A data donation tool to facilitate research on WhatsApp. *Mobile Media & Communication* (2024), 20501579251326809.

[11] Kiran Garimella, Princessa Cintaqia, Juan José Rojas-Constain, Bharat Kumar Nayak, and Aditya Vashistha. 2025. Global Patterns of Viral Content on WhatsApp. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 19. 586–601.

[12] Kiran Garimella and Dean Eckles. 2020. Images and misinformation in political groups: Evidence from WhatsApp in India. *Harvard Kennedy School Misinformation Review* (2020).

[13] Sharad Goel, Duncan J Watts, and Daniel G Goldstein. 2012. The structure of online diffusion networks. In *Proceedings of the 13th ACM conference on electronic commerce*. 623–638.

[14] Mark S Granovetter. 1973. The strength of weak ties. *American journal of sociology* 78, 6 (1973), 1360–1380.

[15] Natalie-Anne Hall, Brendan Lawson, Cristian Vaccari, and Andrew Chadwick. 2023. Beyond quick fixes: How users make sense of misinformation warnings on personal messaging. (2023).

[16] Antonis Kalogeropoulos and Patrícia Rossini. 2025. Unraveling WhatsApp group dynamics to understand the threat of misinformation in messaging apps. *New Media & Society* 27, 3 (2025), 1625–1650.

[17] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 137–146.

[18] Yehuda Lindell. 2020. Secure multiparty computation. *Commun. ACM* 64, 1 (2020), 86–96.

[19] CSDS Lokniti. 2018. How widespread is WhatsApp's usage in India? https://www.livemint.com/Technology/O6DLmIibCCV5luEG9XuJWL/How-widespread-is-WhatsApps-usage-in-India.html

[20] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[21] Katherine L Milkman and Jonah Berger. 2012. What makes online content viral. *Journal of marketing research* 49, 2 (2012), 192–205.

[22] Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, and Rasmus Kleis Nielsen. 2019. Reuters Institute Digital News Report 2019 . Reuters Institute for the Study of Journalism.

[23] United Kingdom Home Office. 2023. End-to-end encryption and child safety — gov.uk. https://www.gov.uk/government/publications/end-to-end-encryption-and-child-safety/end-to-end-encryption-and-child-safety. [Accessed 08-10-2025].

[24] Billy Perrigo. 2019. How WhatsApp Is Fueling Fake News Ahead of India's Elections. *TIME* (2019). https://time.com/5512032/whatsapp-india-election-2019/ Accessed: 2025-10-01.

[25] 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.

[26] Mariana Rosenblat, Inga Trauthig, and Samuel Woolley. 2024. Covert Campaigns: Safeguarding Encrypted Messaging Platforms from Voter Manipulation. *NYU Stern Center for Business and Human Rights Report* (2024).

[27] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *science* 311, 5762 (2006), 854–856.

[28] Connie Moon Sehat, Tarunima Prabhakar, and Aleksei Kaminski. 2021. Ethical approaches to closed messaging research: considerations in democratic contexts. In *MisinfoCon Elections Conference Proceedings*.

[29] Gerry Shih. 2023. Inside the vast digital campaign by Hindu nationalists to inflame India. *The Washington Post* (2023). https://www.washingtonpost.com/world/2023/09/26/hindu-nationalist-social-media-hate-campaign/ Accessed: 2025-10-01.

[30] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon Kleinberg. 2012. Structural diversity in social contagion. *Proceedings of the national academy of sciences* 109, 16 (2012), 5962–5966.

[31] Duncan J Watts and Peter Sheridan Dodds. 2007. Influentials, networks, and public opinion formation. *Journal of consumer research* 34, 4 (2007), 441–458.

[32] WhatsApp. 2019. About forwarding limits | WhatsApp Help Center — faq.whatsapp.com. https://faq.whatsapp.com/1053543185312573. [Accessed 07-10-2025].

A Robustness checks

To validate the robustness of our findings from Section 5.5.2, we conducted a comprehensive sensitivity analysis across five key model parameters, comparing the performance of the Independent Cascade (IC) and Linear Threshold (LT) models on the WhatsApp network. The outcome metric reported in all experiments is the fraction of nodes activated, calculated as the mean number of activated nodes divided by the total number of nodes in the network. This fraction represents the reach or cascade size achieved by each model under different parameter configurations.

Table 2 shows the robustness of our simulation results in Section 5.5.2 for various parameters.

Network Size. We tested the models on networks ranging from 5,000 to 50,000 nodes. Across this tenfold increase in scale, the IC model consistently achieved approximately 24% reach, dropping to only 15.8% on the sparsest 50,000-node configuration where fewer eligible hub nodes were available. In contrast, the LT model maintained approximately 90% reach regardless of network size. This stark difference reveals that simply increasing the network size does not help IC overcome its fundamental limitation: each edge activation is governed by a single independent probability. LT, however, continues to activate the vast majority of the giant component because it accumulates reinforcing signals from multiple neighbors.

Monte Carlo Simulation Runs. Varying the number of simulation runs from 50 to 500 had no effect on the average outcomes, serving only to tighten confidence intervals. IC consistently achieved 25.4 ± 0.1% reach, while LT maintained 88.8 ± 0.2% reach.

Seed Set Size. The two models responded dramatically differently to changes in the initial seed set. IC proved remarkably insensitive to seed size: even with just 50 initial adopters, it achieved approximately 25% reach, nearly identical to configurations with 500 seeds. LT, by contrast, was highly sensitive to the seed set, with reach climbing from 38% (50 seeds) through 69%, 78%, and 85%, finally reaching 89% with 500 seeds. This finding directly reflects the core mechanism of complex contagion: without a sufficiently large initial pool of adopters to create reinforcing exposures in clustered neighborhoods, LT cascades stall. Once enough seeds are present to trigger reinforcement effects, adoption saturates through the network.

IC Transmission Probability. We varied the edge activation probability in the IC model from 0.05 to 0.20. As expected, IC reach

increased linearly with this parameter, rising from 17% to 33%. However, even at the highest transmission probability tested, IC failed to approach LT's performance. Meanwhile, LT remained stable at approximately 89% across all IC probability values, because LT adoption is governed by threshold-based reinforcement rather than single-edge success rates.

LT Threshold Distribution. We explored different θ distribution parameters for the LT model's adoption thresholds. When thresholds were skewed toward easier adoption using $\theta(0.7, 1.3)$, LT reached 95% of the network. Even with $\theta(3.0, 3.0)$—which assigns higher thresholds making adoption more difficult—LT still activated approximately 93% of nodes. Throughout these variations, IC remained unchanged at roughly 25%, further demonstrating that reinforcement mechanisms, not single-edge transmission strength, drive large-scale cascades in this network.

These sensitivity analyses reveal that the performance gap between IC and LT is structural and robust across parameter choices. The IC model's single-exposure mechanism fundamentally limits it to reaching about one-quarter of the network unless transmission probabilities are set unrealistically high or the network topology is substantially altered. The LT model's reinforcement mechanism, by contrast, aligns naturally with the structure of the content-augmented WhatsApp network: once users observe multiple neighbors adopting content, their thresholds are crossed and cascades propagate through nearly the entire giant component. Combined with the dose-response analysis showing the sigmoidal adoption curve, these findings provide converging evidence that content virality in this WhatsApp dataset exhibits the hallmarks of complex contagion rather than simple independent cascade dynamics.

Table 2: Sensitivity Analysis of Model Parameters

Network Size		
Nodes	IC (%)	LT (%)
5,000	24.5	87.9
10,000	25.5	89.0
20,000	23.8	90.5
30,000	23.4	90.8
50,000	15.8	88.9
Monte Carlo Steps		
Runs	IC (%)	LT (%)
50	25.5	88.8
100	25.5	89.0
200	25.4	88.8
300	25.4	89.0
500	25.4	88.9
Seed Counts		
Seeds	IC (%)	LT (%)
50	25.5	38.5
150	25.4	69.3
250	25.5	78.4
350	25.5	85.3
500	25.5	89.1
IC Activation Probability		
p	IC (%)	LT (%)
0.05	16.8	89.0
0.08	21.0	88.6
0.12	25.5	88.8
0.15	28.4	88.8
0.20	32.8	89.2
LT Threshold Distribution		
(α, β)	IC (%)	LT (%)
(0.7, 1.3)	25.4	95.4
(1.0, 1.0)	25.5	88.8
(1.5, 1.5)	25.4	91.4
(2.0, 2.0)	25.4	92.7
(3.0, 3.0)	25.4	93.0