Informative Integrity Frictions in Social Networks

Lluís Garcia-Pueyo
Facebook
Menlo Park, USA
lgp@fb.com

Samantha Guthrie
Facebook
Menlo Park, USA
samguthrie@fb.com

Bernardo Santana Schwarz
Facebook
Menlo Park, USA
bsantana@fb.com

Baoxuan Xu
Facebook
Menlo Park, USA
baoxuanxu@fb.com

ABSTRACT

Social media platforms such as Facebook and Twitter benefited from massive adoption in the last decade, and in turn facilitated the possibility of spreading harmful content, including false and misleading information. Some of these contents get massive distribution through user actions such as sharing, to a point that content removal or distribution reduction does not always stop its viral spread. At the same time, social media platforms efforts to implement solutions to preserve its integrity are typically not transparent, causing that users are not aware of any integrity intervention happening on the site. In this paper we present the rationale for adding what are now visible friction mechanisms to content share actions in the Facebook News Feed, its design and implementation challenges, and results obtained when applying them in the platform. We discuss effectiveness metrics for such interventions, and show their effects in terms of positive integrity outcomes, as well as in terms of bringing awareness to users about potentially making harmful content viral.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**; • **Information systems** → **Social networks**.

KEYWORDS

social network, integrity, machine learning, share friction, viral, misinformation

ACM Reference Format:
Lluís Garcia-Pueyo, Samantha Guthrie, Bernardo Santana Schwarz, and Baoxuan Xu. 2022. Informative Integrity Frictions in Social Networks. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3487553.3524221

1 INTRODUCTION

Social sites like Facebook and Twitter experienced massive adoption in the last decade, becoming relevant content distribution and information channels. These sites facilitated the distribution and consumption of many types of content, including harmful and low-quality content. As a result, social media platforms have been in need of implementing solutions to measure, detect, and reduce the distribution of such harmful content [5] to preserve their integrity.

Distribution of content in social sites works as follows: a user publishes a piece of content[1], and the content is made available to *viewer* members within the social network of the publisher user. *Viewer* users have in turn the possibility to share[2] the content further, making it available to their social network and followers. This cycle may repeat indefinitely, and the content distributed may as a result reach more users in an exponential pattern. The rapid spread of content using these sharing features may contribute to turn it *viral*.

Content integrity problems are diverse in terms of harm, including (1) content that violates what social networks consider as acceptable[3], (2) content that while accepted produces bad experiences to users due to its low-quality, and (3) borderline content that is not violating but produces bad experiences to some particular users. Integrity work consists in detecting content in these categories, and apply enforcement actions depending on the type of harm being produced. For example, social media cites may remove content that does not adhere to its Community Standards. As another example, low-quality content such as click-bait may be demoted in ranking to reduce its distribution.

Besides these types of harms, there is a class of sensitive content that does not fit the categories above, but whose nature is correlated with integrity harms. Examples of sensitive content are unconfirmed misinformation, out of date articles, highly reported content, etc. These contents may not be classified as an integrity problem, but their distribution might represent a risk if ultimately there's a confirmation that they can be harmful. In addition, such content applied to high risk areas such as COVID and vaccinations topics can represent a true risk for users.

In this paper, we present an Informative Re-Share Friction mechanism that warns users before sharing particular types of sensitive content, and informs them about the risks of sharing that content. This mechanism consists on an interstitial with an informative message adapted to the type of content being shared, and prompting

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France.

ACM ISBN 978-1-4503-9130-6/22/04.
https://doi.org/10.1145/3487553.3524221

[1] Different platforms would name the action as "post", "tweet", or "pin" depending on the site the action was done
[2] "share", "retweet", etc; depending on the platform
[3] Example categories of violating content can be found in the Facebook Community Standards [5]

the user to confirm that the share action should continue. The goals of such mechanism are to help users identify bad content and learn more about the relevant pieces of information related to the content, and to reduce the propagation of integrity harms that benefit from viral distribution. This mechanism has been deployed at the Facebook News Feed, and we present empirical data showing its benefits.

2 RELATED WORK

Our work is directly related to the reduction of harmful content in social networks, user controls in social networks, and sharing behaviors in social networks. There's extensive work in detection of harmful content in social media ranking from Misinformation [12], Toxicity [14], Adult and Graphic imagery [4], and others [8]. Reduction of harmful content is typically focused on content-level understanding and ranking modification which is not transparent to users in social media platforms. In our case, we aim to reduce the distribution of harmful content by being transparent and inform users about the potential harm. Most of the controls work in the space has been dedicated to providing user privacy controls [7][2][9], and we provide additional references to warnings and to provided users with harm information in Section 3 [3][15][13]. Our work is based on the outcomes of these research.

3 USER PROBLEM

Drawing from multiple internal research studies, both qualitative and quantitative in nature, we identified three central people problems, that at first glance may appear in conflict:

(1) Users want social media companies to do more about integrity harms.
(2) Users want to feel more empowered and more in control of their social media experience.
(3) Users want the opportunity to assess information for themselves, but they also want to have the relevant information to do so.

Further, from extensive qualitative work, particularly in low-literacy communities, researchers were aware that some users faced challenges assessing information and identifying low-quality content. As social media platforms were made simpler and sleeker, these challenges only increased. Thus, there's a need to communicate more directly and obviously about potential harm. A simple front-end intervention that empowered users and gave them insight into how and what we considered harmful would be extremely beneficial and likely high in user value.

This approach also aligned well with the external literature, both from misinformation and integrity scholars, as well as the security engineering community. There, researchers state *an effective physical warning clearly communicates risk, consequences of not complying, and instructions to comply (although some of this information can be omitted if the risk is obvious or the consequences can be deduced from the warning)* [3]. Although in the case of re-share interstitials, it's not about compliance, but about conveying the potential risk, be it small (sharing out-of-date articles) or large (sharing confirmed misinformation). The same internal research shows that users do not want to share certain integrity harms, like misinformation and misleading content, and an interstitial is an effective way to provide them with information on how to avoid doing so.

But, social media faces an added complexity: digital risks are often hard to understand. A sign warning of a broken sidewalk is easy to understand [3]. The risk is *open and obvious* [15]. A sign about a potential financial scam, less so. Users are often not familiar with the terminology or complex technical "jargon" such as phrases like "domain age," they do not have the complete context, and cannot "see" the potential risk in the same way they could see a broken sidewalk. This is, of course, sometimes by design - scams work because they look like websites and people you trust. On social media, the risk is even more complex as the scam, or other harm, may be posted by a friend or Page users trust, reducing the level of skepticism users approach the situation with to start.

This complexity puts the onus on social media platforms for how to best communicate the potential risk, help people overcome cognitive biases, make decisions more carefully, and assist users to make the decision they would make if they had complete information. We know that even a small hint to be more discerning can change outcomes [13].

Further, when writing about usability and trust in download warnings, [10] concludes that to keep user trust in warnings high, the only case where a warning is justified is *where genuine danger has been detected within, but there is still a chance that the file in question is not malicious.*

With both the internal and external literature taken together, it becomes evident that warnings are a useful tool in our fight to reduce integrity harms on social media platforms, while providing users with information to help empower them to make more informed decisions in the future.

4 IMPLEMENTATION

The regular Facebook News Feed sharing flow consists on a "Composer" window that appears after the user presses the "Share" button, and that allows the user to add text on top of the content being shared before sharing it with their social network. We will refer to this "Composer" window as the *regular Facebook News Feed sharing flow* moving forward.

The informative integrity friction presented in this paper consists on an interstitial that is shown after a user clicks on the "Share" button on a content of the Facebook News Feed, as seen in Figure 1. The interstitial consists on an informative message, a "Continue" button, and a "Go back" button.

The informative friction is only enabled for certain content types in the Facebook News Feed, which we refer to as *target posts*, and that are identified as suitable for showing the Informative Friction, as explained in Section 3. For example, a *target post* believed to contain information to COVID-content[4], would be suitable for showing the friction. Other examples of content types that would trigger the friction can be seen in Table 1.

The content type of the *target post* determines the text shown in the interstitial. For example, a COVID-related content type would show the text *This post mentions COVID-19 For more info and resources, go to the COVID-19 Information Center. See Info*, while a

[4]As determined by a classifier, or text matching system.

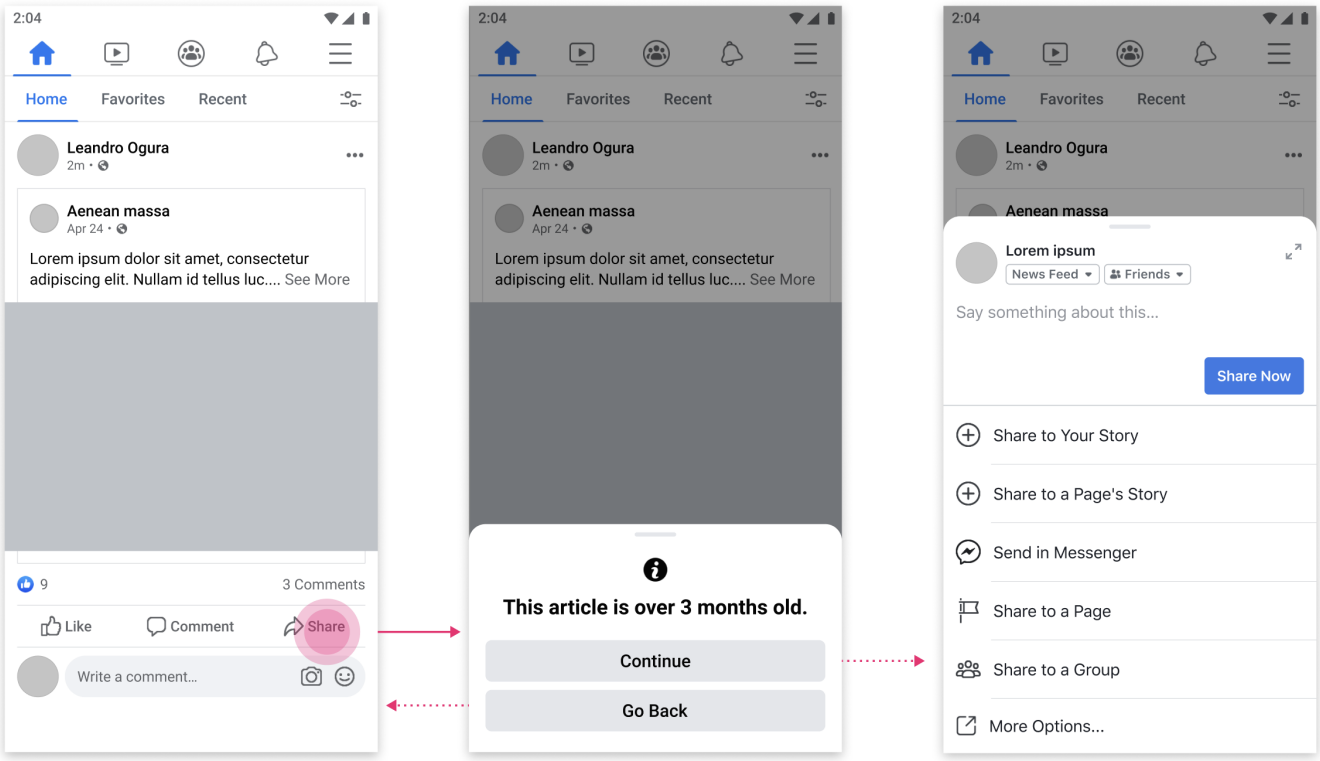

Figure 1: Re-share friction on an article that is more than 3 months old

Content Type	Product Description
COVID	Provides additional context on COVID and directs user to Facebook's information hub
Vaccine	Shows up on vaccine-related content and provides additional information about vaccinations
Outdated News	Appears on News posts that are older than 90 days
Confirmed Misinformation	Shows up on articles that have been fact-checked by third-party fact-checkers

Table 1: Example Re-share Interstitials

Vaccine related content type would show the text *This post mentions vaccines For more information about vaccines, visit cdc.gov.*.

Once the interstitial is shown, if the user clicks the "Continue" button, the user continues to the regular Facebook News Feed sharing flow which allows for the *target post* to be shared to the users' network. If the user clicks "Go back" or clicks outside of the interstitial, the friction is closed, and the sharing flow is cancelled. Figure 1 shows an example workflow of Re-share friction for post with content that is more than 3 months old.

The visual appearance of the interstitial differs depending on the platform shown. Thus, for example, while in mobile devices

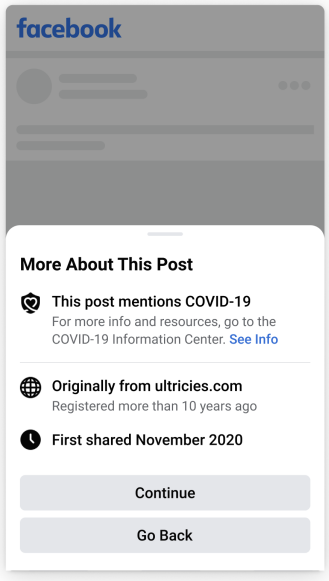

Figure 2: COVID Re-share friction on mobile.

is shown in the bottom-sheet style shown in Figure 2, in desktop browsers is shown in the pop-up style in Figure 3.

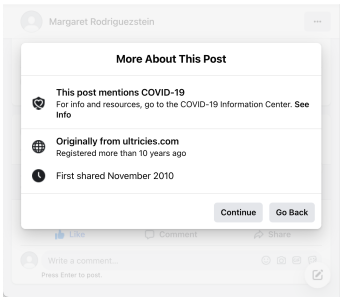

Figure 3: COVID Re-share friction on desktop.

5 EXPERIMENTS

To validate the effectiveness of the mechanism, we run a viewer-level A/B test that measures the causal effect of this user interface change in the Facebook News Feed. The change consists on the interstitial presented in Section 4 that appears once a user attempts to share a link that mentions COVID-related content. The user interface contains information and links that direct the user to Facebook's COVID hub site[5], and also provides some basic information about the website that hosts the link to be shared, as shown in Figure 2. Content that is eligible for this experiment was determined by a multilingual regular expression detecting COVID-related content with an accuracy of 98% [11]. The hypothesis is that given the additional context, users will reconsider sharing links that might be false or misleading, thus reducing the total consumption of misinformation in the News Feed.

We define two equally sized groups for our experiment. Users in the treatment group did observe the interstitial when sharing links that are classified as COVID-related, and are presented with two options: continue sharing or going back (cancelling). Users in the control group did not see the interstitial and continued into the regular Facebook News Feed sharing flow.

We compare the two groups on two main metrics: total number of views on confirmed misinformation (i.e. posts that have been marked as misinformation by third-party fact-checkers [6][1]), and the interstitial *cancel rate*, which is defined as the ratio of times the sharing process was cancelled due to pressing the "Go back" button (defined as *explicit cancel rate*), or by click outside of the interstitial (defined as *implicit cancel rate*). All numbers reported are statistically significant at the 95% level and we report the point estimate.

The experiment shows an overall *cancel rate* of 45.82% and an *explicit cancel rate* of 4.51%. We also observe the *habituation* of users to newly added controls in Figure 4, since there's typically a novelty effect which makes feature usage decrease over time. In this case, the *cancel rate* decreases and flattens over time and stabilizes at around the fifth view. Among those who have seen the interstitial at least ten times, cancel rates decrease 25pp from the first to the tenth interstitial experience. The stabilization after multiple exposures would prove the utility of the friction even after user *habituation.*

As shown in Table 2, views on confirmed misinformation decrease 5.25% for general topics and 12.90% for health-related links.

[5]https://www.facebook.com/coronavirus_info/

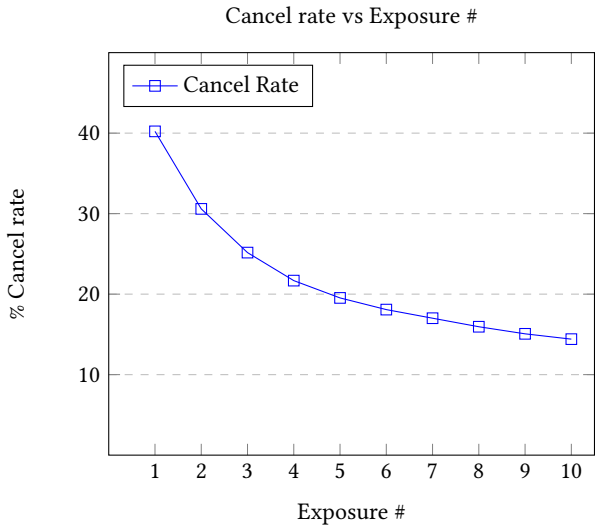

Figure 4: Habituation in users with at least 10 exposures to the interstitial

Confirmed Misinformation Prevalence		
	Overall Views	Views on Confirmed Misinformation
All links	-0.44%	-5.25%
Health links	-1.53%	-12.90%

Table 2: Prevalence of fact-checked Misinformation

The reduction of overall misinformation validates the hypothesis stated above. Note that the results are remarkable given they correspond to a viewer-level A/B test and the effects of adding friction elements to share actions would show its full potential to downstream viewers (those potentially seeing the content being shared). We plan on exploring producer-side effects of this treatment in future work.

6 CONCLUSION & NEXT STEPS

In this paper, we presented an Informative Integrity Friction that was deployed in the Facebook News Feed, and that is triggered when users share content types that are related to integrity harms. The presented mechanism aims to better inform users about the content that is shared in social networks, as well as to prevent the viral spread of integrity harms in such networks. Experimental results validate that the proposed approach indeed contributes to reduce the overall prevalence of harmful content in social networks. Experimental results also show that the feature continues being useful after habituation, which further motivates its long term usage. As next steps, we plan on exploring other mechanisms to provide with more controls and contextual information to users about integrity harms.

REFERENCES

[1] Product Management Anna Stepanov, Director. 2021. Sharing Our Content Distribution Guidelines. Facebook Newsroom. (Sep 2021). https://about.fb.com/news/2021/09/content-distribution-guidelines/

[2] Pauline Anthonysamy, Awais Rashid, and Phil Greenwood. 2011. Do the Privacy Policies Reflect the Privacy Controls on Social Networks?. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. 1155–1158. https://doi.org/10.1109/PASSAT/SocialCom.2011.150

[3] Cristian Bravo-Lillo, Lorrie Faith Cranor, Julie Downs, and Saranga Komanduri. 2011. Bridging the Gap in Computer Security Warnings: A Mental Model Approach. *IEEE Security Privacy* 9, 2 (2011), 18–26. https://doi.org/10.1109/MSP.2010.198

[4] Thomas M Chen. 2021. Automated Content Classification in Social Media Platforms. In *Securing Social Networks in Cyberspace*. CRC Press, 53–71.

[5] Facebook. [n.d.]. Facebook Community Standards. Facebook Transparency Center. ([n. d.]). https://transparency.fb.com/policies/community-standards/?from=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2F

[6] Facebook. [n.d.]. Facebook's Third-Party Fact-Checking Program. Facebook Transparency Center. ([n. d.]). https://www.facebook.com/journalismproject/programs/third-party-fact-checking

[7] Ryan Galpin and Stephen V. Flowerday. 2011. Online social networks: Enhancing user trust through effective controls and identity management. In *2011 Information Security for South Africa*. 1–8. https://doi.org/10.1109/ISSA.2011.6027520

[8] Alon Y. Halevy, Cristian Canton-Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, and Ves Stoyanov. 2020. Preserving Integrity in Online Social Networks. *CoRR* abs/2009.10311 (2020). arXiv:2009.10311 https://arxiv.org/abs/2009.10311

[9] Irina Heimbach and Oliver Hinz. 2018. The Impact of Sharing Mechanism Design on Content Sharing in Online Social Networks. *Information Systems Research* 29, 3 (2018), 592–611. https://doi.org/10.1287/isre.2017.0738 arXiv:https://doi.org/10.1287/isre.2017.0738

[10] Kat Krol, Matthew Moroz, and M Angela Sasse. 2012. Don't work. Can't work? Why it's time to rethink security warnings. In *2012 7th international conference on risks and security of internet and systems (CRiSIS)*. IEEE, 1–8.

[11] Igor L. Markov, Jacqueline Liu, and Adam Vagner. 2021. Regular Expressions for Fast-response COVID-19 Text Classification. *CoRR* abs/2102.09507 (2021). arXiv:2102.09507 https://arxiv.org/abs/2102.09507

[12] Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabbany. 2021. The Surprising Performance of Simple Baselines for Misinformation Detection. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 3432–3441. https://doi.org/10.1145/3442381.3450111

[13] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science* 31, 7 (2020), 770–780.

[14] Martin Saveski, Brandon Roy, and Deb Roy. 2021. The Structure of Toxic Conversations on Twitter. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1086–1097. https://doi.org/10.1145/3442381.3449861

[15] Michael S Wogalter. 2006. *Handbook of warnings*. CRC Press.