9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Anticipated versus Actual Effects of Platform Design Change: A Case Study of Twitter's Character Limit

Keywords: platform design, Twitter, user behavior, predictability, demographics

Extended Abstract

Introduction. Online platform design is, on the one hand, critically important, given that the online content can reach and affect people worldwide, and, on the other hand, very challenging, given the intrinsic entanglement between the digital environment and user behavior. The design of platforms impacts user behavior [2], and user behavior, in turn, shapes the platforms [4]. It is challenging to predict how users will respond to design changes and new platform policies as any changes to large complex socio-technical systems may lead to unintended consequences.

Here we examine an instance of a particularly important real-world platform design change: on 7 November 2017, Twitter suddenly and unexpectedly increased the maximum allowed tweet length from 140 to 280 characters, thus altering its signature feature. According to Twitter, this change, which we henceforth refer to as "the switch", was introduced to give users more space to express their thoughts, as a disproportionately large fraction of tweets had been exactly 140 characters long, reflecting users' need to "cram" or "squeeze" their tweets.

The design of this platform change and the choice of the new policy was informed by modeling historical user behavior [3]. When deciding how to design the platform change, Twitter engineers and policy-makers made estimates and predictions based on data from before the intervention. In particular, to estimate users' need to "cram" their tweets, they estimated the number of tweets impacted by the policy, that is, the number of tweets that would be longer than the character limit if they could be longer. The new policy, 280 characters, was selected since it was estimated that, if that limit were enforced [3], a negligibly low fraction of tweets would afterwards be impacted by cramming.

Twitter's decision to double the maximum allowed tweet length is, therefore, a remarkable example of a platform design intervention that was informed by modeling based on publicly available historical user traces. However, it is known that there are factors that render anticipating the impact of real-world interventions challenging [1]. Nonetheless, Twitter engineers and policy-makers, by their own account [3], fitted a model to historical user behavior traces and implicitly assumed that the user behavior would not change in response to the implementation of the change. Such a static, "no-response" view might or might not hold. Is it necessary to account for user response, or does the actual user behavior remain faithful to what was anticipated?

Methods. In our analysis, we contrast Twitter's anticipated pre-intervention predictions about user behavior with actual post-intervention user behavior: Did the platform design change lead to the intended user behavior shifts, or did a gap between anticipated and actual behavior emerge? Did different user groups react differently? Applying the same approach as Twitter policy-makers, we create updated counterfactual estimates and find whether the character limit would need to be increased further to reduce cramming that re-emerged at the new limit. Using a 1% sample of all tweets spanning the period from 1 January 2017 to 31 October 2019, we model tweet length over time (illustrated in Fig. 1).

Results. We find that gaps between anticipated and actual user behavior emerged after the intervention. Initially, the estimated amount of cramming at 140 characters was aligned with actual user behavior. However, actual user behavior eventually diverged from anticipated user behavior. While cramming at 140 characters sharply decreased after the introduction of 280 characters, cramming, although less drastic, shifted to the new length limit. Furthermore, examining tweet text, syntactic and semantic indicators also provide evidence of cramming that emerged at the new length limit. Overall, these gaps between anticipated and actual user behavior are more pronounced on the Web interface compared to mobile devices.

Furthermore, we consider hypothetical interventions that would reduce the cramming that emerged post-intervention. We find that as user behavior shifts in response to a platform change, the estimated effects of hypothetical policies change as well (Fig. 2). Given that 280 characters were selected to make a vast majority of tweets fit the limit, post-intervention, since the cramming re-emerged, the necessary number of characters would have to be increased further to achieve the same objective (Fig. 3).

Discussion. We contribute to the rich literature studying online user behavior with an empirical study that reveals a dynamic interaction between platform design and user behavior. Our results emphasize a dynamic interaction between platform design and user behaviors. The length limit was doubled in order to reduce users' need to cram their tweets. However, the intervention did not entirely solve the issue, as cramming re-emerged at the new length limit. Before the intervention, the modeled data was "collected under the policy", with character limit shaping the data that Twitter subsequently based their measurements on. When modeling in such a static regime, the validity of estimates is threatened, and it is complicated to evaluate alternative policies before their deployment. As the new policy is deployed, new behaviors and new data are collected under a different policy which elicits behaviors that are not anticipated—although can be explained after the fact. This fusion of platform design decisions and user behaviors can lead to feedback loops and calls for more cautious approaches that take into account the dynamic nature of user response. These findings highlight the fluidity of online behaviors and have direct implications for large-scale user behavior studies, human–computer interaction, and platform design.

References

[1] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proc. ACM Hum.-Comput. Interact. (CSCW)*, 2, 2018.

[2] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences (PNAS)*, 119(1), 2022.

[3] Ihara Ikuhiro. Our discovery of cramming, 2017. `https://web.archive.org/web/20220428070306/https://blog.twitter.com/engineering/en_us/topics/insights/2017/Our-Discovery-of-Cramming`.

[4] Claudia Wagner, Markus Strohmaier, Alexandra Olteanu, Emre Kıcıman, Noshir Contractor, and Tina Eliassi-Rad. Measuring algorithmically infused societies. *Nature*, 595(7866):197–204, 2021.

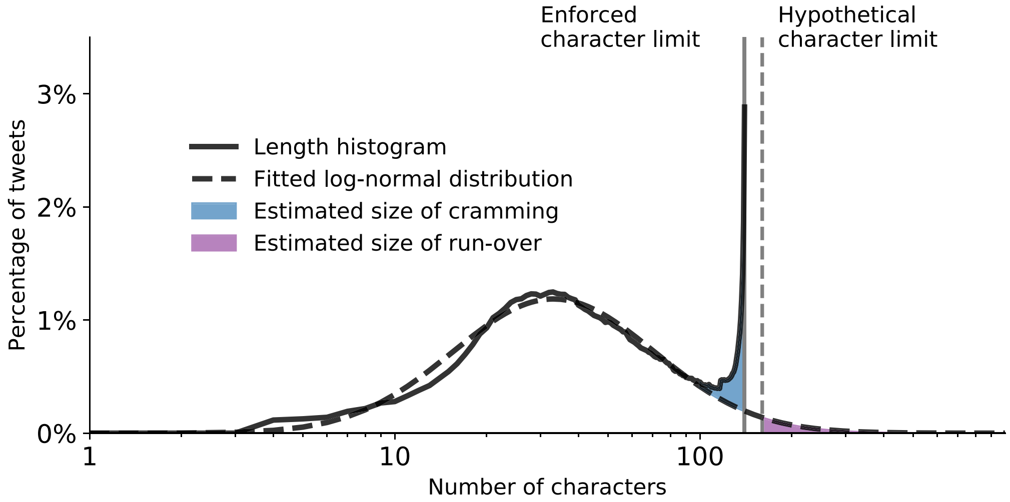

Figure 1: **Illustration of model for measuring impact of character limit.** The example depicts the length distribution of English tweets across mobile and Web devices, before the 280-character limit was introduced. The solid black line shows the length histogram; the dashed black line marks the fitted log-normal distribution (logarithmic x-axis). The solid gray vertical line marks the enforced character limit (140 characters in the example). The dashed gray vertical line marks a hypothetical character limit (150 characters in the example). *Cramming* (marked as the blue area) is the deviation of the empirical distribution from the underlying log-normal distribution near the character limit. *Run-over* (marked as the purple area) is the part of the fitted log-normal distribution that falls beyond the given number of characters. Cramming can only be used to measure people's attempts to "squeeze" their tweets within the *enforced character limit* (solid vertical line). On the contrary, run-over is used to evaluate the impact of *any hypothetical character limit* (dashed vertical line).

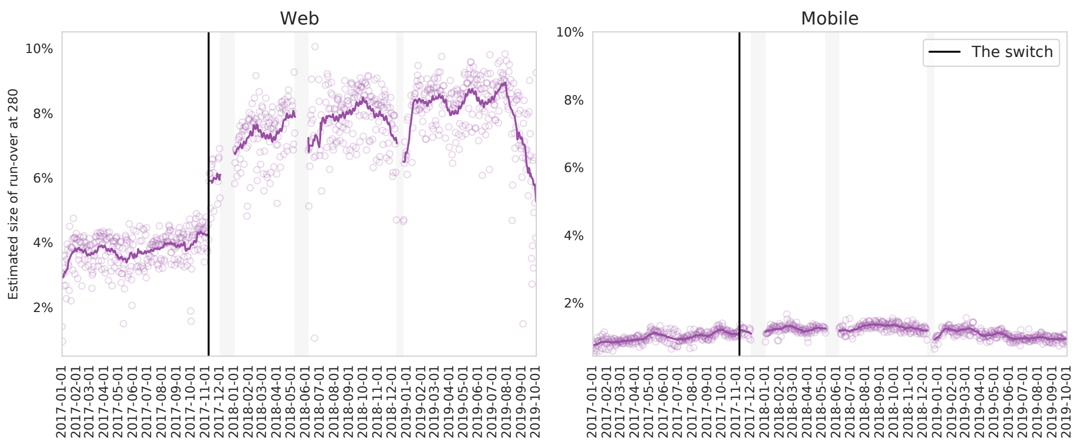

Figure 2: **Evolution of estimated size of run-over at 280 characters.** Circles indicate the daily estimated fraction of tweets that would be longer than 280 characters if it were possible; the line marks the 10-day rolling average. The quantities are shown separately for the Web interface (left) and mobile applications (right). The vertical line marks the switch, and gray bands mark days with missing data. After the switch, the estimated size of run-over at 280 characters increases sharply on the Web interface, while it increases only slightly on mobile devices.

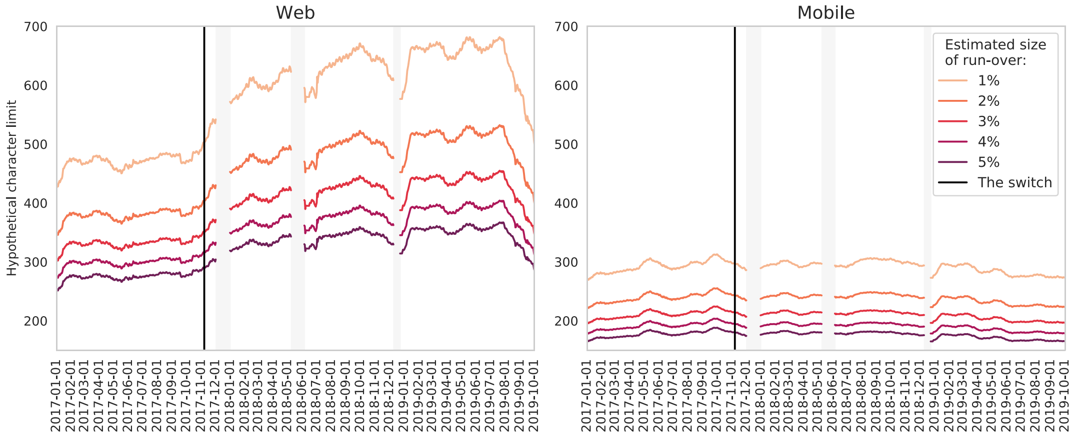

Figure 3: **Evolution of hypothetical character limit necessary to achieve various targeted sizes of run-over.** Tweet length limits (in the number of characters) are shown on the y-axis; targeted sizes of run-over are marked in colors. Lines mark 10-day rolling averages. The quantities are shown separately for Web interface (left) and mobile applications (right). The vertical line marks the switch and gray bands mark days with missing data. After the switch, the number of characters needed to achieve a targeted run-over increases sharply on the Web interface, while it remains robust on the mobile devices.