Investigating How Gilds Were Employed On Reddit

Charlotte Lambert
cjl8@illinois.edu
Computer Science
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Yoshee Jain
yosheej2@illinois.edu
Computer Science
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Koustuv Saha
ksaha2@illinois.edu
Computer Science
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Eshwar Chandrasekharan
eshwar@illinois.edu
Computer Science
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Abstract

Certain types of positive feedback are among the most commonly-used signals in online communities. For the recipients, these signals may already be serving as positive reinforcement, a psychology principle effective at encouraging desired behaviors in offline settings. Furthermore, our preliminary work shows that Reddit moderators are explicitly providing positive feedback to reinforce behavior. Even though users and moderators may currently be reinforcing behavior through positive feedback, we do not understand what type of content is being reinforced. We aim to fill that gap by uncovering how often prosocial metrics from prior work appear in posts that receive positive feedback compared to those that do not. We find that existing measures of prosocial behavior are insufficient for capturing the content-level differences between rewarded posts and non-rewarded posts. We call for future work to better understand this problem and to dive deeper into the area of positive reinforcement in moderation.

CCS Concepts

• **Human-centered computing → Empirical studies in collaborative and social computing**.

Keywords

Social Computing; Positive Feedback; Prosocial Behavior.

ACM Reference Format:
Charlotte Lambert, Yoshee Jain, Koustuv Saha, and Eshwar Chandrasekharan. 2024. Investigating How Gilds Were Employed On Reddit. In *Companion of the 2024 Computer-Supported Cooperative Work and Social Computing (CSCW Companion '24), November 9–13, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3678884.3681916

CSCW Companion '24, November 9–13, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-1114-5/24/11
https://doi.org/10.1145/3678884.3681916

1 Introduction

"We heard you... awards are back!"– Reddit Product Team[1]

In September 2023, gilds and other positive feedback mechanisms were removed from Reddit entirely[2] as the platform worked to create improved ways of empowering communities to reward content. Although *gilding*—i.e., the ability to "gild" posts by donating a Reddit Premium subscription—is no longer a feature on Reddit, the platform brought back awards in May 2024[1] based on feedback from its users. This recent attempt at reinventing the way users can provide positive feedback demonstrates the importance of such mechanisms, both from the perspective of the platform and the users themselves. We take advantage of the fact that so many stakeholders are invested in the availability and usage of positive feedback tools to motivate our work exploring how exactly users were engaging with one award mechanism in particular: gilds.

In this research, we turn to communities themselves to identify what constitutes high quality contributions through their use of awards—in particular, gilds. This parallels prior research in the space of content removals which inherently trusts members of Reddit communities to provide us with a ground truth to study undesirable behavior [6, 13]. While removals are a mechanism only available to moderators, we highlight awards like gilds, a form of positive feedback that was available to all members of a community.

1.1 Research Questions

Although other forms of awards have been brought back, Reddit has not revived its gild feature. The platform's users were displeased that gilds—one of Reddit's most beloved features—was not returning [1]. Given users' fondness for gilds in particular, our goal is to investigate how they were employed before Reddit removed this functionality in 2023.

Specifically, we ask the following research questions:

RQ1: How often did Reddit communities give out gilds?
RQ2: What types of content received gilds in Reddit communities?

To answer these research questions, we collect a dataset of posts and summarize the usage of gilds in the dataset, highlighting the rarity with which they are awarded. Then, in order to understand

[1] https://www.reddit.com/r/reddit/comments/1css0ws/we_heard_you_awards_are_back/
[2] https://www.reddit.com/r/reddit/comments/14ytp7s/reworking_awarding_changes_to_awards_coins_and/

Table 1: This table describes the minimum, mean, maximum, and total values of various summary statistics across the subreddits in our dataset. Out of 1,024,066 subreddits that posted within our study period, 13,815 (1.3%) subreddits awarded at least one gild.

	Minimum	Mean	Maximum	Total
# Posts	1	63.10	1,064,270	64,621,215
# Gilded Posts	0	0.14	4,957	145,954
# Unique Gilded Authors	0	0.12	1,978	119,469
Mean Gilds per Post	0	0.01	12	-

the type of content that received gilds, we identify a group of treatment users (i.e., users who received a gild on a post) and a control group of similar users who did not receive the treatment. This allows us to use existing measures of prosocial behavior and toxicity from prior work to compare the content of the posts and identify any observable differences.

2 Background

Although positive feedback (e.g., upvotes/likes, awards, etc.) is a substantial aspect of many users' experiences in online communities, there are many avenues of social computing research left to explore before we understand what type of content is being rewarded and what the impact of receiving such feedback is on users. Additionally, positive feedback may be a powerful tool for community moderation by allowing communities themselves to reinforce desired behaviors. Research has previously explored platforms that engage with community moderation practices [15], however there are platforms like Reddit which do not explicitly identify themselves as community-moderated that may already be utilizing such practices implicitly.

In our preliminary work, we surveyed moderators to understand the role of positive reinforcement in their moderation practices [12]. We found that many moderators explicitly provide positive feedback as a way to reinforce behavior they want to encourage, including prosocial behavior, quality, and engagement with other users. The survey also uncovered specific actions moderators take to reinforce in practice. These actions included many signals in the Reddit interface, some of which are specific to moderators and others that are widely accessible. We highlight one method used by moderators that was available to all users: gilds.

Gilds were an award-like signal users can give to posts and consisted of gold, silver, and platinum icons. They were a way for users to purchase coins to award to posts, making them a potentially strong signal of quality given the rarity with which they were given out. Gilds, among other positive feedback mechanisms on Reddit, can be considered examples of gift-giving in an online context. Social computing literature establishes the importance of gift-giving in a community as users feel a sense of reciprocity (i.e., they will receive similar treatment by the community in the future) [11]. Gift-giving also builds reputation, another important concept in social computing literature shown to be an incentive for good behavior [17], a way to clearly signal desired behaviors [14], and a mechanism for users to build trust and encourage engagement [18].

Other studies have shown that gift-giving not only encourages reciprocity for the giver, but also for the recipient. More specifically, gift-giving increases the likelihood that the recipient will give a similar gift in the future [10]. If positive feedback mechanisms, such as gilds, abide by this principle, receiving positive feedback on posts may encourage the recipients to give out more positive feedback in the future. Thus it is important for us to study the types of content getting rewarded as well as the outcomes for users receiving such positive feedback from the community.

3 RQ1: How often did Reddit communities give out gilds?

We construct a dataset, $\mathcal{D}$, using Pushshift's historical Reddit archives [4] of posts between May 1, 2020 to September 30, 2020 (our study period) from all sub-communities (i.e., subreddits) on the platform. Table 1 includes a summary of $\mathcal{D}$.

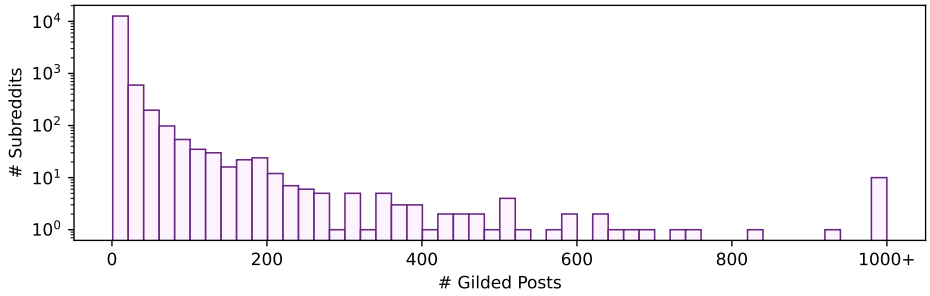
From this summary table, we see that posts are rarely gilded. Out of the 64.6M posts in $\mathcal{D}$, only 145K (0.2%) received at least one gild from its community. This is largely impacted by the fact that only 1.3% of the subreddits in $\mathcal{D}$ actually engage with gilds.

Figure 1a visualizes the distribution of gilded posts per subreddit. This only includes the subreddits that have at least one gilded post in the dataset. We see that the vast majority of subreddits that gild do so infrequently. The most common number of gilded posts for a subreddit to have within our four month study period is 1.

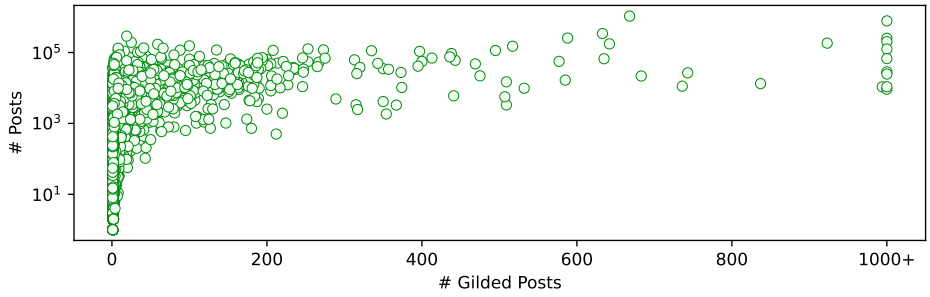
However, there are quite a few communities which used gilds substantially more frequently, with some even gilding hundreds (or thousands) of posts in the study period.

We also plot the number of gilded posts in each subreddit against the total number of posts made in the subreddit during our study period (Figure 1b). This plot demonstrates the wide range in subreddit activity levels in the set of subreddits that utilize gilds. There are many subreddits with fewer than 10 posts in the four month window, but also many with hundreds of thousands of posts. This plot also shows that even in subreddits with similar posting frequency habits, some gilded nearly zero posts in the study period while others gilded upwards of 1000. Finally, there are some subreddits that gilded a substantial percentage of their posts. For example, r/RedditSessions, a subreddit ranked in the top 1% of subreddits by size, gilded 4.9K out of 24.7K (20.1%) of the posts in our study period.

This descriptive analysis demonstrates that gilds, while rare, were still an important part of many Reddit communities and are deserving of further analysis. We also argue that the rarity may support the idea that gilds are high-quality signals of desirable behavior on Reddit.



(a) This histogram reports the distribution of the number of gilded posts across subreddits.



(b) This scatter plot shows the number of gilded posts compared against the total number of posts.

Figure 1: These figures show data only from the subreddits in $\mathcal{D}$ that have at least one gilded post. Subreddits with more than 1,000 gilded posts are binned together.

4 RQ2: What types of content received gilds on Reddit?

To explore what posts were getting positive feedback in the form of gilds, we focus specifically on the subset of posts from $\mathcal{D}$ posted in the 100 subreddits which use gilds most frequently.

4.1 Constructing Treatment and Control groups

First, we identify a set of treated candidates as the posts in $\mathcal{D}$ which received at least one gild. We take a sub-sample of the remaining posts to create a pool of control candidates three times as large as our pool of treatment candidates.

Next, we perform within-subreddit stratified matching on propensity scores [8, 16, 19–22] to minimize the differences in pre-treatment author and environmental characteristics between our control and treatment groups. This method aims to minimize biases present in propensity score matching identified by King and Nielsen [9]. Our propensity scores are based on several covariates related to pre-treatment activity, author-level characteristics, and post-level characteristics.

The resulting propensity score reflects how likely a post is to belong to the treatment group based on the given covariates. Within each subreddit, we bin the propensity scores into 10 strata using quantiles, each representing a group of posts with a similar likelihood of being treated. Thus, the result is a set of many-to-many matches for each subreddit such that every treated post in a stratum is matched with all control posts in the stratum.

4.2 Metrics

In order to compare the content of treated posts against the content of control posts, we look specifically at the title and body of posts in our treated and control datasets. In the case that a post has no text body (i.e., the post contains only images, videos, or links), we only use its title.

Here we describe the metrics applied to the post content. For each treatment, we report the mean value of each metric in the Treated and Control columns of Table 2.

4.2.1 Lexicon-Based Counts. Bao et al. [3] identify a handful of metrics for capturing prosocial behavior, including three lexicon-based counts that we utilize for our content analysis. First, we count instances of *laughter* in each post using a lexicon of laughter words (e.g., "haha"). Second, we count how much *gratitude* appears in a post using a lexicon of words and phrases such as "thank you". Finally, we count the number of *donation* instances in each post by looking at fundraising URLs.

4.2.2 BERT Models. Using models provided by Bao et al. [3], we calculate scores for *politeness*, *support*, and *agreement* for posts in our treatment and control groups. The models output a score between 1 and 5 where 5 represents strong evidence of politeness, support, or agreement and 1 represents the negative extreme.

4.2.3 Measures of Toxicity. Since it is also possible that toxicity affects whether a post receives positive feedback from its community, we calculate toxicity scores for posts using models trained by Almerekhi et al. [2]. For each post, the model outputs probabilities

Table 2: This table reports the mean values of each metric for the Treated and Control groups. We also report t-tests and Cohen's d. Values in the t-test column with * have $p < 0.05$, ** have $p < 0.01$, and * have $p < 0.001$.**

Metric	Gild			
	Treated	Control	t-test	Cohen's d
Laughter	0.01	0.01	-0.96	-0.02
Gratitude	0.03	0.01	10.35***	0.18
Donation	0.00	0.0	1.81	0.03
Politeness	3.12	3.07	8.93***	0.16
Support	3.07	3.03	7.64***	0.13
Agreement	2.95	2.92	7.33***	0.13
Highly Toxic	0.05	0.06	-3.44***	-0.06
Slightly Toxic	0.17	0.19	-4.92***	-0.09
Non-Toxic	0.77	0.75	5.41***	0.09

between 0 and 1 that the post is highly toxic, slightly toxic, and non-toxic. For each post, these three probabilities sum to 1.

4.3 Analysis

We conduct two-sided independent t-tests between the treatment and control group to determine whether treated users have significantly higher or lower values of each metric. We report the t-statistic in Table 2 and indicate statistical significance. We also calculate Cohen's d [7], a measure that reflects the size of the difference between the means of the treatment and control groups, to gain more insight into the practical significance of our findings. According to Cohen [7], values of 0.2, 0.5, and 0.8 correspond to small, medium, and large effect sizes respectively.

4.4 Findings

As shown in Table 2, treated posts have significantly more instances of gratitude and significantly higher politeness, support, and agreement scores compared to control posts. Table 2 also highlights that treated posts are less toxic on average than control posts. These findings demonstrate that many aspects of prosocial behavior are found more frequently in posts that received gilds.

Despite the observed statistically-significant differences for many of our metrics, the Cohen's d values are all below 0.2, the established threshold for small effects [7]. As a result, we conclude that the differences in the metrics we utilized have limited practical significance and the effects observed are very small.

5 Discussion and Future Work

In this section, we discuss the findings from our analysis of gilded posts and propose future directions to further our understanding of positive feedback mechanisms on Reddit.

5.1 The Implications of Removing Gilds

Our prior survey found that moderators across different subreddits agreed that prosocial behavior should be encouraged in their communities [12]. The survey also found that identifying high-quality content to encourage is a challenge moderators face given the focus of moderation tools on punitive actions (e.g., content removal, user bans, etc.). In this work, we found that gratitude, politeness, support, agreement, and non-toxicity all appeared significantly more often in posts awarded gilds by the users of a community than in non-gilded posts. As a result, before being removed by the platform, gilds may have been an easy way to identify prosocial posts, and thus posts that moderators would want to encourage. In other words, Reddit recently removed a valuable signal of prosocial behavior that could have been used by moderators in their efforts to positively reinforce high-quality content. Platform designers should make sure there are useful signals available to be used by moderators to facilitate the identification of high-quality content. Future work is needed to understand whether other existing signals are similarly capable of highlighting prosocial content as gilds.

5.2 Develop More Robust Measures of Quality

Our research utilizes metrics grounded in prior work to make progress toward understanding what types of posts received positive feedback in the form of gilds. We demonstrated that these metrics are not capable of capturing the difference between posts that received positive feedback and posts that did not. As a result, this research demonstrates a need for future work towards understanding what qualities of a post increase the likelihood of it receiving positive feedback. Specifically, we call for methods of understanding rewarded content in a community-specific way to allow for variations between subreddit norms and perspectives. This may involve an analysis similar to the approach taken by Chandrasekharan et al. [6] to learn about overlaps in norms through cross-community similarities in content removals, but from a standpoint of identifying desirable content.

5.3 Other Forms of Positive Feedback

Additionally, this work focuses on gilds which are just one form of awards historically available through Reddit's interface. Furthermore, gilds were a paid feature, thus they were given out infrequently. While this may indicate that gilds are a strong signal of quality and desirability in a community, the cost may also have

been a deterrent for some users to engage with the feature. As a result, it is possible that gilds were disproportionately given out by users from higher socio-economic backgrounds, introducing potential bias in the type of content that is getting rewarded. Given that gilds are not a feature on Reddit anymore, future work is needed to explore content being rewarded with other forms of positive feedback, especially awards that are free to give out and still available through Reddit's interface.

5.4 Identify the Causal Effects of Receiving Awards

Finally, this initial exploration into content that receives positive feedback motivates the need for research to understand the causal effect of receiving such feedback on the recipients. We are currently working on a project that will explore the user-level effects of receiving forms of positive feedback on their Reddit posts. This project utilizes the dataset we developed for RQ2 and involves a difference-in-differences analysis similar to prior work [5] to understand how receiving a gild impacted several user-level outcomes (e.g., participation, quality of contributions, etc.). We find evidence that receiving a gild encouraged users to make contributions more likely to receive positive feedback from other users. This is promising initial evidence that forms of positive feedback, such as gilds, may be able to encourage higher-quality contributions.

This future work has the potential to inform platforms and communities about how they can utilize positive reinforcement mechanisms to support distributed moderation practices and improve contributed content on a user level.

6 Conclusion

From our exploration of the usage of gilds on Reddit, we found that this type of award was sparsely used, both on a community- and post-level. However, many subreddits of varying sizes and activity levels utilized gilds, motivating the need for future investigation into the role of gilds in positive reinforcement.

In our content analysis of gilded posts, we found that existing measures of prosociality and toxicity are insufficient to capture the differences between posts that received gilds and those that do not. This result indicates that we need more sophisticated measures to capture the differences between these two groups of posts.

Overall, our work demonstrates a gap in the current social computing landscape: measures of desirability. We also established many future avenues to explore, specifically related to other forms of positive feedback (e.g., awards) and the causal effects of receiving positive feedback for individual users.

References

[1] [n. d.]. Reddit brings back its old award system — 'we messed up' - The Verge. https://www.theverge.com/2024/5/17/24158848/reddit-brings-back-award-system-gold-coins-messed-up

[2] Hind Almerekhi, Haewoon Kwak, and Bernard J. Jansen. 2022. Investigating toxicity changes of cross-community redditors from 2 billion posts and comments. *PeerJ Computer Science* 8 (Aug. 2022), e1059. https://doi.org/10.7717/peerj-cs.1059

[3] Jiajun Bao, Junjie Wu, Yiming Zhang, Eshwar Chandrasekharan, and David Jurgens. 2021. Conversations Gone Alright: Quantifying and Predicting Prosocial Outcomes in Online Conversations. In *Proceedings of the Web Conference 2021*. ACM, Ljubljana Slovenia, 1134–1145. https://doi.org/10.1145/3442381.3450122

[4] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 830–839.

[5] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 1. 1–22. https://doi.org/10.1145/3134666

[6] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 2. 1–25. https://doi.org/10.1145/3274301

[7] Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences* (2. ed., reprint ed.). Psychology Press, New York, NY.

[8] Emre Kiciman, Scott Counts, and Melissa Gasser. 2018. Using Longitudinal Social Media Analysis to Understand the Effects of Early College Alcohol Use. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (June 2018). https://doi.org/10.1609/icwsm.v12i1.15012 Number: 1.

[9] Gary King and Richard Nielsen. 2019. Why Propensity Scores Should Not Be Used for Matching. *Political Analysis* 27, 4 (Oct. 2019), 435–454. https://doi.org/10.1017/pan.2019.11

[10] René F. Kizilcec, Eytan Bakshy, Dean Eckles, and Moira Burke. 2018. Social Influence and Reciprocity in Online Gift Giving. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–11. https://doi.org/10.1145/3173574.3173700

[11] Peter Kollock et al. 1999. The Economies of Online Cooperation: Gifts and Public Goods in Cyberspace. *Communities in cyberspace* 239 (1999).

[12] Charlotte Lambert, Fred Choi, and Eshwar Chandrasekharan. 2024. "Positive reinforcement helps breed positive behavior": Moderator Perspectives on Encouraging Desirable Behavior. *Proceedings of the ACM on Human-Computer Interaction* CSCW (2024).

[13] Charlotte Lambert, Ananya Rajagopal, and Eshwar Chandrasekharan. 2022. Conversational Resilience: Quantifying and Predicting Conversational Outcomes Following Adverse Events. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 548–559. https://doi.org/10.1609/icwsm.v16i1.19314

[14] Cliff Lampe. 2012. The role of reputation systems in managing online communities. *H. Masum, M. Tovey, eds* (2012), 77–88.

[15] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 543–550.

[16] Alexandra Olteanu, Onur Varol, and Emre Kiciman. 2017. Distilling the Outcomes of Personal Experiences: A Propensity-scored Analysis of Social Media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 370–386. https://doi.org/10.1145/2998181.2998353

[17] Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. Reputation systems. *Commun. ACM* 43, 12 (2000), 45–48. Publisher: ACM New York, NY, USA.

[18] Paul Resnick, Richard Zeckhauser, John Swanson, and Kate Lockwood. 2006. The value of reputation on eBay: A controlled experiment. *Experimental economics* 9 (2006), 79–101. Publisher: Springer.

[19] Koustuv Saha and Amit Sharma. 2020. Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), 590–601. https://doi.org/10.1609/icwsm.v14i1.7326

[20] Koustuv Saha, Ingmar Weber, and Munmun De Choudhury. 2018. A Social Media Based Examination of the Effects of Counseling Recommendations after Student Deaths on College Campuses. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (June 2018). https://doi.org/10.1609/icwsm.v12i1.15016

[21] Gaurav Verma, Ankur Bhardwaj, Talayeh Aledavood, Munmun De Choudhury, and Srijan Kumar. 2022. Examining the impact of sharing COVID-19 misinformation online on mental health. *Scientific Reports* 12, 1 (May 2022), 8045. https://doi.org/10.1038/s41598-022-11488-y

[22] Yunhao Yuan, Koustuv Saha, Barbara Keller, Erkki Tapio Isometsä, and Talayeh Aledavood. 2023. Mental Health Coping Stories on Social Media: A Causal-Inference Study of Papageno Effect. In *Proceedings of the ACM Web Conference 2023*. ACM, Austin TX USA, 2677–2685. https://doi.org/10.1145/3543507.3583350