Reviews in Motion: A Large Scale, Longitudinal Study of Review Recommendations on Yelp

Ryan Amos*, Roland Maio†, Prateek Mittal*, Joseph A. Calandrino‡
*Princeton University, †Columbia University, ‡Federal Trade Commission
rbamos@cs.princeton.edu, rjm2212@columbia.edu, pmittal@princeton.edu, jcalandrino@ftc.gov

***Abstract*—The United Nations Consumer Protection Guidelines list "access ... to adequate information ... to make informed choices" as a core consumer protection right. However, problematic online reviews and errors by algorithms that detect those reviews pose obstacles to the fulfillment of this right. Research on problematic reviews and defense algorithms often derives insights from a single web crawl, but the filtering decisions those crawls observe may not be static: these algorithms may cause a review to disappear or reappear after the crawl. These often-opaque changes may frustrate not only authors whose reviews vanish but also consumers and affected businesses. For example, a business may go from a 3.5- to 4.5-star rating despite receiving no new reviews. We introduce a novel longitudinal angle to the study of problematic reviews. We focus on "reclassification," wherein a platform changes its filtering decision for a review. We performed repeated web crawls of Yelp to create three longitudinal datasets. These datasets highlight the platform's dynamic treatment of reviews. We compiled more than 2M unique reviews across over 10K businesses in 12.5M data points. Our datasets are available for researchers to use.**

Our longitudinal approach gives us a unique perspective on Yelp's classifier and allows us to explore reclassification. We find that more than 8% of reviews may move between Yelp's two main classifier classes ("Recommended" and "Not Recommended"), raising concerns about prior work's use of Yelp's classes as ground truth. Some reviews move multiple times: we observed up to five reclassifications in eleven months. Our data suggests demographic disparities in reclassifications, with more changes in lower income and low-middle density areas.

I. Introduction

Online reviews are a key source of consumer information, play an important role in consumer protection, and have a substantial impact on businesses' economic outcomes [1]–[3]. This creates incentives for various parties to engage in problematic reviewing practices [4], [5]. These practices are of ongoing interest to regulators and encompass a large spectrum of behaviors, such as creation of new accounts for posting fake reviews, hijacking of legitimate accounts, compensation of real users for posting favorable reviews, incentivized reviews generally, reviews written by friends or competitors, hiding of negative reviews, requests for negative reviews be handled confidentially, or reviews that are not relevant to the product or service [6]–[9]. To address the challenges problematic reviews pose, review platforms have created classification systems to present users with the best reviews to make informed decisions. A wealth of prior work has taken a variety of approaches to understanding online reviews and these classification systems. This work ranges from studying the factors motivating both honest and problematic reviews to exploring how to detect and generate opinion spam or fake reviews [6], [10], [11]. Most of this work examines reviews and review classification systems at a single point in time. Because classification decisions are not static, this perspective is incomplete.

This report was prepared in part by staff of the Federal Trade Commission's Office of Technology Research and Investigation and does not necessarily reflect the views of the Commission or any individual Commissioner.

Research questions. We sought to address the following research questions:

1) How frequent is review reclassification on Yelp?
2) How do factors such as review age, author, and geography correlate with reclassification on Yelp?

Within geographic factors, we focused on population density and median household income.

Contributions. In this work, we move from a view of reviews at rest to a view of reviews in motion. We explore the dynamic classification of reviews through the collection and analysis of three novel, longitudinal datasets, focused on Yelp. We sought to observe the changing nature of review classification, and thus we collected datasets in which we observed reviews for a fixed set of businesses at multiple times. We focus on the movement of reviews between Yelp's "Recommended" and "Not Recommended" classifications, which we call "reclassification."

Our contributions are as follows:

1) The largest longitudinal dataset of reviews. We collected over 2M unique reviews of more than 10K businesses. We monitored those reviews over periods from four months to eight years, resulting in 12.5M data points.
2) The first study of review reclassification, including an in-depth examination of this previously unexplored phenomenon. We find reclassification rates—the percentage of reviews reclassified during the study period—from 0.54% over 4 months to 8.69% over 8 years. We uncover demographic disparities in reclassification.

Implications. Our results demonstrate that reviews are routinely reclassified, occasionally multiple times. These reclassifications suggest details of Yelp's classifier, the degree of confidence in classifications, the difficulty of classifying reviews, and more. The frequency of reclassification also raises questions regarding studies that depend on Yelp's classifications as ground truth [12]–[17]. Reviews reclassified

multiple times—of which we observed 1,233 cases—suggest that reclassifications do not always move towards ground truth. Balancing fairness to legitimate reviewers with the need to remove problematic reviews is a challenging problem, and our work sheds light on these challenges.

We provide an extended version of this work and access to our dataset at https://sites.google.com/princeton.edu/longitudinal-review-data/.

II. Related work

Extensive academic literature from multiple disciplines studies online reviews and fake reviews. We highlight four primary areas of prior work: longitudinal study of reviews, demographics and reviews, problematic reviews, and incentives for reviewing. To the best of our knowledge, no other studies focus on review reclassification. Yelp acknowledges that an automated system classifies and sometimes reclassifies reviews, but Yelp does not disclose the frequency with which reclassification happens [7], [18].

Longitudinal study. Some researchers have performed longitudinal analyses on reviews with a single snapshot, for example by using review dates [19]–[21]. Other studies have linked datasets together to gain a better view of the review landscape, but still rely on a single snapshot for each data point [22]. Our work provides the first perspective on how review classifications change.

Demographics and reviews. Ensuring equal access to and treatment by technology is an important equity issue. Prior work explores how regional factors like population density impact review posting frequency [11], [19], [23], [24], with mixed and sometimes conflicting findings. We offer additional details on the interaction between demographics and reviews.

Problematic reviews. Problematic reviews have been a persistent challenge in the review landscape, particularly detecting fake reviews. Ott et al. estimate that fake reviews occurred at a rate of 2-6% across six platforms [25], but other estimates of fake reviews reach 50%-70% [26], [27]. Yelp reports filtering about one quarter of its reviews [28]. A wealth of research focuses on detecting fake reviews [6], [14], [16], [20], [29]–[31]. Some have taken an adversarial approach, generating fake reviews rather than detecting them [17], [32], [33].

One challenge in analyzing fake reviews is the absence of ground-truth data. Fake reviews may be designed to fool even humans [34]. To overcome this challenge, prior work has taken several approaches, such as using leaked data from fake reviewers [35], posing as customers to fake review providers to identify fake reviews [29], employing hand-crafted heuristics like duplication [6], paying study participants to create artificial fake reviews [34], and using Yelp's classifications [12], [14]. Our work highlights a danger in using Yelp's labels as ground truth for a problematic review classifier.

III. Data collection

We collected and constructed three longitudinal datasets to study reviews on Yelp.

A. Background

Yelp divides reviews into two primary categories, assigned by a software classifier. The categories are "Recommended" and "Not Recommended." Yelp lists four considerations in classifying a review as "Not Recommended:" conflicts of interest, solicited reviews, reliability, and usefulness. Not Recommended reviews do not affect metrics like a business's average star rating, and Yelp displays these reviews less prominently [7], [36], [37]. Yelp also has a third review class—"Removed for Violating our Terms of Service"—which we do not use in our analysis. While Yelp has published an official review dataset for academic purposes [38], this dataset is not up-to-date and includes only Recommended reviews.

We chose to study Yelp because prior work had established reference datasets we could compare against [14]. Furthermore, Yelp allows access to reviews that it does not recommend, unlike many other platforms.

B. Datasets

We describe each dataset that we compiled below. Table I provides summary details and statistics for these datasets.

Coarse Dataset (EYG). The first dataset includes a single recrawl of the same businesses that Mukherjee et al. [14] previously crawled, enabling us to observe reclassifications of the reviews. We call this the "eight year gap (EYG)" dataset because Mukherjee et al. performed their crawl in 2012 and we performed ours in 2020.

Fine Dataset (CHI). To examine reclassification in finer detail, we built our second dataset by collecting reviews from businesses in a given set of ZIP Codes repeatedly: eight times over eleven months. To provide some overlap with the EYG businesses, we included the same ZIP Codes from the EYG dataset. These ZIP Codes are all in the Chicago metropolitan area, so we call this the "Chicago (CHI)" dataset. This more comprehensive but localized coverage of reviews allows us to observe things like multiple reviews from the same authors.

Population Density and Income (UDIS / UDS & UIS). Our third dataset (UDIS) is really a pair of datasets that offers a more representative range of reviews across the US: our "US Density Stratified (UDS)" and "US Income Stratified (UIS)" datasets. We stratified US regions along two axes: population density and median household income. We collected density and income information from the US Census [39]. We used ZIP Code Tabulated Areas (ZCTAs) as a proxy for ZIP Code. ZCTAs approximate and typically match USPS ZIP Codes [40].

For the density dataset, we divided ZIP Codes into five strata by population [41]. We uniformly sampled ZIP Codes from each strata until we had sampled at least 500 businesses from that strata, using the Yelp Fusion API to determine how many businesses were in each ZIP Code. We collected four monthly crawls for each business. We repeated the same process for the income dataset.

The strata for the income crawls are: \$0–\$55k, \$55k–\$68k, \$68k–\$82k, \$82k–\$105k, \$105k–\$250k; for density: 0–67, 67–302, 302–881, 881–1,873, 1,873–57,541 ppl/km^2.

TABLE I
COMPOSITION OF THE DATASETS.

	EYG	CHI	UDS	UIS
Timespan	8 years	11 months	4 months	4 months
# Reviews	263,308	10,485,007	1,409,059	1,145,995
# Unique reviews	196,383	1,395,870	358,184	292,107
# Businesses	201	5,773	2,829	2,843
# Crawls	2	8	4	4
# Authors (range)	100,713 - 119,037	404,706 - 520,195	212,348 - 259,862	180,994 - 221,591
% Reclassified	8.69%	0.87%	0.54%	0.61%
% Recommended	88.19%	88.90%	85.69%	85.22%

C. Collection

For each dataset, collection occurred in two phases. First, we had an initial setup phase to compile the set of businesses to crawl. Next, we crawled for and collected reviews one or more times. At the time of each crawl, we visited and collected all reviews for each business. We provide an overview of our collection process in Appendix B, Figure 7, and a list of data and metadata collected in Appendix B, Table IV.

Business data. To collect data from Yelp, we first identified the businesses to crawl. For the EYG dataset, we used Yelp's Fusion API [42] to collect business URLs for the business identifiers we had. For the CHI and UDIS datasets, we used Yelp's Fusion API for each targeted ZIP Code to collect a list of all businesses. In situations where the search exceeded the API's response limit, we divided our query using other search parameters to reduce the size of the response. We took the union of the businesses returned by all queries and excluded any results that did not have an address with a targeted ZIP Code. For each dataset, we determined our targeted business list initially, and it remained static over all crawls.

Crawling procedure. We used Pyppeteer [43] in headless mode for our web crawler. To mitigate IP blocking, we used a VPN. For each crawl, we iterated over each ZIP Code, then each business in it. We navigated to the business's page, then each Recommended review page. If we received a block page or exceeded 100 page loads since we changed our VPN connection, we connected to a new VPN server. We then navigated to the Not Recommended reviews, where we repeated the above process. The CHI and UDIS crawls were repeated multiple times for a longitudinal perspective. Appendix B, Figure 6 shows the timeline of the crawls. Appendix A details additional quality control and post-processing steps.

Ethics. We identify two sources of ethical concerns: the privacy of the user data we collected, and the impact of our collection on Yelp's servers. While all data collected is, or was, publicly available, review authors did not agree to have their data included. We treat fields such as author name and review text as sensitive. We require researchers requesting sensitive data to provide an adequate justification. To minimize the impact of our research on Yelp's servers, we limited the number of simultaneous crawling threads to six, throttled our crawlers, and built our crawlers to minimize the pages scraped.

TABLE II
RECLASSIFICATION OF REVIEWS BETWEEN 2012 AND 2020 (E.G. TOP RIGHT IS REVIEWS NOT RECOMMENDED IN 2012 BUT RECOMMENDED IN 2020). INCLUDES ONLY REVIEWS PRESENT IN BOTH SNAPSHOTS.

	Rec. (2012)	Not Rec. (2012)
Rec. (2020)	56,048	3,566
Not Rec. (2020)	2,249	5,059

IV. RESULTS

Our findings hint at details of Yelp's approach to classification, possible changes to its classifier, confidence in classifications, and the difficulty of the problem. They also suggest the impact reclassifications may have on reviewers, consumers, and businesses. For studies that use Yelp's classifications as ground truth [12]–[17], our observations may help in assessing the validity of particular findings.

In the long run, reviews disproportionately migrate to Recommended. We first consider reclassification over the long term with the coarse EYG dataset. Table II shows how review classes changed between the two crawls. Reclassification is not uncommon: although most reviews remain in the same class, a significant number are classified differently between crawls (χ^2 test with 1 degree of freedom: $p << 1e-5$). Despite the fact that 87.6% of the reviews were already Recommended in the 2012 crawl, more reviews are reclassified as Recommended from Not Recommended than vice versa. Proportionately, 3.9% of reviews that were Recommended in the 2012 crawl were Not Recommended in 2020, but 41.3% of reviews that were Not Recommended in 2012 were Recommended in 2020. This suggests that Yelp errs towards the Not Recommended class initially for reviews and corrects later.

Many reviews are reclassified; a few are reclassified frequently. To investigate the frequency and scale of reclassification in the short term, we use the finer CHI dataset. Table III shows how frequently reviews were reclassified. Approximately 0.8% of reviews were reclassified in this dataset. Despite the shorter study period and limited number of crawls, a small fraction of reviews underwent a substantial number of changes. A few reviews changed classes almost every measurement. These reviews may be cases that are particularly hard for Yelp to classify.

In the long run, newer Not Recommended reviews are

TABLE III
RECLASSIFICATION PATTERNS IN CHI DATASET OVER EIGHT CRAWLS. "R" DENOTES RECOMMENDED, "N" DENOTES NOT RECOMMENDED.

# changes	Pattern	Count
0	R	1,235,194
	N	148,278
	Total	1,383,472
1	R → N	4,953
	N → R	5,573
	Total	10,526
2	R → N → R	706
	N → R → N	373
	Total	1079
3+	R → N → R → N	60
	N → R → N → R	75
	R → N → R → N → R	14
	N → R → N → R → N	15
	N → R → N → R → N → R	2
	Total	157

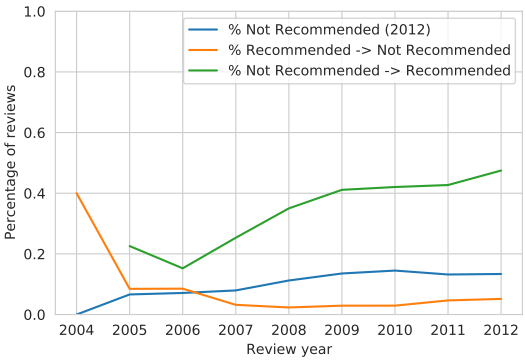

Fig. 1. Change in recommendation status by the year reviews posted. The blue line represents reviews Not Recommended in 2012; the orange and green lines represent reviews that were Recommended and Not Recommended, respectively, in 2012 but reclassified in 2020.

more likely to be reclassified. To explore the relationship between reclassification and review age, we group reviews in EYG by the year posted. Within each group, we measure the percentage of reviews Yelp recommended in 2012 as well as the percentage of 2012 Recommended and Not Recommended reviews reclassified in 2020 (Figure 1). Note that 2004 has just five reviews, and all were Recommended in 2012. The trend in percentage Not Recommended in 2012 that are Recommended in 2020 (green line) suggests that Yelp was more likely to reclassify newer reviews from Not Recommended to Recommended. Yelp could have had more time to examine older reviews by 2012, or older, less sophisticated problematic reviews may have resulted in fewer errors to correct. Alternatively, this could stem from the use of review age or correlated factors (e.g., the number of reviews by the author) in classification. Yelp suggests that it considers how established a reviewer is when recommending reviews [7], [36].

Newer reviews are more likely to undergo repeated reclassification, but the chance persists over time. Figure 2 shows that reviews which undergo more frequent reclassification in our fine CHI dataset tend to be newer.[1] This supports the idea that Yelp increases its confidence in classifying reviews as the reviews age, perhaps because Yelp has observed more activity from the author's account or the business. Nevertheless, we see reclassification of reviews dating back to 2005, so a review's classification may never be fully stable.

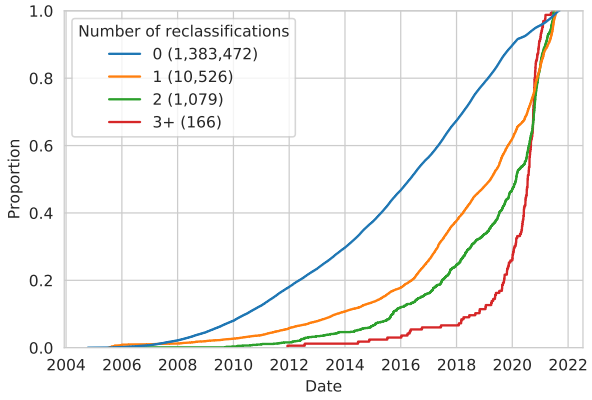

Fig. 2. Cumulative percentage of reviews with a given number of reclassifications that were posted by a given date. For example, by the start of 2018 approximately 20% of reviews with 2 observed reclassifications had been posted.

If newer reviews are more likely to undergo reclassification, the average reclassification rate should decrease as reviews grow older. The EYG and CHI creation processes (and Figure 2) suggest EYG reviews are older, but the rates for EYG (8.69% over 8 years; 0.09%/month) and CHI (0.87% over 11 months; 0.08%/month) are similar. The cause is unclear: for example, Yelp's classifier may be more stable today, or Yelp could have overhauled its classifier between the EYG crawls.

Review classes follow the author. To determine if reclassifications are performed at the author level, we investigated whether authors who have a review classification change are likely to have their other reviews match the new classification. In the 1,175 cases in which an author with multiple reviews in the CHI dataset had a review reclassified, 924 had all of their reviews match after reclassification. The average percentage of an author's review pairs that matched classification was 95.1% for Recommended reviews and 94.7% for Not Recommended reviews. This suggests that classifications follow the author.

Demographic factors correlate with reclassification frequency. To test whether Yelp reclassifies reviews for businesses in regions with certain income or density attributes more frequently, we looked at the average number of reclassifications per review in each stratum for both the UIS and UDS crawls (Figure 3). Less dense and lower income regions experience more reclassification. The 60-80% strata are an outlier in both cases—the 60-80% density stratum experiences substantially more reclassifications while the 60-80% income stratum experiences substantially fewer than its

[1]Note a small artifact in the upper-right corner of Figure 2: the 1, 2, and 3+ lines rise above the 0 line. The newest reviews cannot have 1, 2, or 3+ reclassifications, since we have fewer observations.

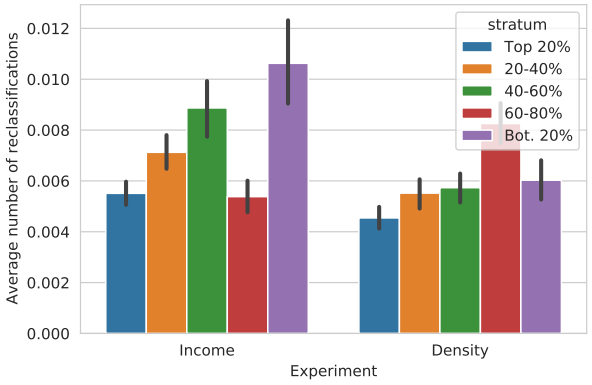

Fig. 3. Average number of reclassifications per unique review per stratrum. Black bars indicate the 95% confidence interval

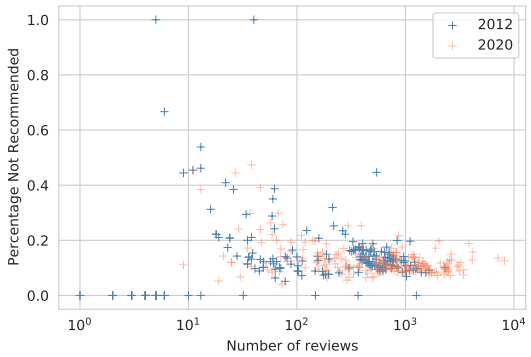

Fig. 5. Number of reviews versus the percentage of reviews that were Not Recommended for a business, 2012 and 2020 data (EYG dataset).

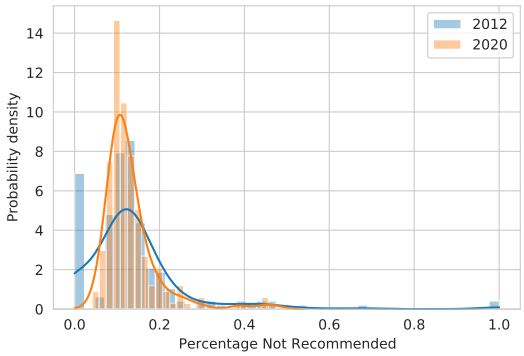

Fig. 4. Probability density of Not Recommended review percentage for a business in 2012 and 2020 (EYG dataset). Lines are kernel density estimates.

neighbors, but on par with the top income stratum, which warrants further research. These disparities invite questions as to why they arise: could something about these regions lead to more challenging-to-classify reviews, might Yelp's classifier not be well tuned to them, etc.?

With more reviews, businesses tend towards a similar percentage of Not Recommended reviews. We used the EYG dataset to understand whether review classification has evolved at the business level. For 2012 and 2020, we examined (1) the distribution of the percentage of Not Recommended reviews per business and (2) the connection between number of reviews and percentage Not Recommended per business. Figure 4 shows the distribution of percentage Not Recommended by business. The median percentage Not Recommended remains similar (0.122 and 0.115), but the distributions are distinct (Kolmogorov–Smirnov test $p < 0.01$). In 2012, many business have no Not Recommended reviews.

These results may be explained in part by businesses having fewer reviews in 2012 than 2020. Figure 5 shows the connection between the number of reviews of a business and the percentage Not Recommended. Most businesses converge to around the same percentage Not Recommended with enough reviews. This convergence appears tighter for the 2020 data. While we did not establish a significant correlation between the number of reviews and the percentage Recommended for the 2012 data (Spearman's correlation $\rho = 0.13$, $p = 0.06$), a significant, negative correlation exists for the 2020 data ($\rho = -0.31$, $p < 1e-5$). In other words, the more reviews a business has, the smaller the proportion of Not Recommended reviews. The correlation for the 2020 data may be more significant because the businesses are more established.

V. Discussion

While we discuss reclassifications largely in the abstract, they may have real consequences for reviewers, consumers, and businesses. One business had its star rating go from 3.5 to 4.5 after a set of reclassifications. Another went from 2.5 to 1. Such swings can have a substantial impact on the businesses and on consumers who place their trust in reviews. In another case, a reviewer had 18 reviews temporarily reclassified as Not Recommended. Yelp may have temporarily obscured a considerable amount of legitimate feedback from this reviewer, or it may be displaying 18 problematic reviews to consumers.

Reviews on Yelp routinely move between classifications, in both directions, occasionally multiple times. Our observations suggest Yelp errs towards classifying newer reviews as Not Recommended—correcting later—and tends to makes decisions on a per-reviewer basis rather than a per-review basis. We also identify potential demographic disparities in reclassification frequency. Our approach can offer insights into reclassification for reviewers, businesses, and consumers, all of whom are impacted by this opaque practice.

Our analysis has lessons for platforms beyond Yelp. Our observations suggest a relatively conservative classification approach by Yelp: until reviewers establish themselves, Yelp trusts their reviews less. This approach has downsides—including potentially obscuring legitimate reviews—but may be effective for other platforms looking to improve review quality. Yelp's practice of obscuring suspected problematic reviews but keeping them accessible is also atypical. This practice provides some transparency and may work well for other platforms.

We observed potential demographic disparities in Yelp's practices. Platforms (and perhaps policymakers) should carefully consider such disparities: what inequities exist, what are their impacts on consumers and businesses, and how can any negative impacts be mitigated? Even seemingly minor features like reclassification may create or exacerbate inequalities.

Limitations. Our study is limited to a single platform and a single country, so it may not be representative. The study period includes the COVID-19 pandemic, a period of substantial social and economic disruption [44], [45]. Local median income and population density may not accurately reflect all businesses and reviewers in a region. We expect that we missed some data, such as reclassifications that occurred between crawls. Finally, Yelp may have removed some problematic reviews for terms of service violations rather than classifying them as Not Recommended.

Future work. Future work could explore additional platforms or further investigate income, density, and other demographic disparities. It could also consider reviews that Yelp removed for terms of service violations, which could include problematic reviews. Finally, it could examine other related practices, like editing and deleting reviews and business data.

Availability. Our crawling and analysis software is available at https://sites.google.com/princeton.edu/longitudinal-review-data/, and our pseudonymized dataset is available for researchers to access. We hope that our work brings momentum to the longitudinal study of review platforms and helps advance consumers' access to quality information.

ACKNOWLEDGMENTS

We thank Amy Winecoff, Anne Kohlbrenner, Ben Kaiser, Dan Salsburg, Hugh Huettner, Laura DeMartino, and Sayash Kapoor for their helpful feedback and suggestions. We thank the anonymous reviewers for their insightful comments.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. (DGE-1644869). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] M. Luca, "Reviews, reputation, and revenue: The case of Yelp.com," *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper*, no. 12-016, 2016.

[2] C. K. Anderson, "The impact of social media on lodging performance," *The Center for Hospitality Research*, 2012.

[3] United Nations, "United Nations guidelines for consumer protection," New York and Geneva, 1985, Revised 2016.

[4] D. Streitfeld, "Buy reviews on Yelp, get black mark," *The New York Times*. [Online]. Available: https://www.nytimes.com/2012/10/18/technology/yelp-tries-to-halt-deceptive-reviews.html

[5] J. R. Miller, "Plastic surgeon, wife used fake Yelp reviews to damage rival: suit," *New York Post*, 2019. [Online]. Available: https://nypost.com/2019/11/07/plastic-surgeon-wife-used-fake-yelp-reviews-to-damage-rival-suit/

[6] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, 2008, pp. 219–230.

[7] Yelp, "Why would a review not be recommended?" [Online; accessed 8-February-2021]. [Online]. Available: https://www.yelp-support.com/article/Why-would-a-review-not-be-recommended

[8] "FTC puts hundreds of businesses on notice about fake reviews and other misleading endorsements," *FTC News*, Oct 2021. [Online]. Available: https://www.ftc.gov/news-events/press-releases/2021/10/ftc-puts-hundreds-businesses-notice-about-fake-reviews-other

[9] "Fashion Nova will pay $4.2 million as part of settlement of FTC allegations it blocked negative reviews of products," *FTC News*, Jan 2022. [Online]. Available: https://www.ftc.gov/news-events/press-releases/2022/01/fashion-nova-will-pay-42-million-part-settlement-ftc-allegations

[10] K. H. Yoo and U. Gretzel, "What motivates consumers to write online travel reviews?" *Information Technology & Tourism*, vol. 10, no. 4, pp. 283–295, 2008.

[11] J. Baginski, D. Sui, and E. J. Malecki, "Exploring the intraurban digital divide using online restaurant reviews: A case study in Franklin County, Ohio," *The Professional Geographer*, vol. 66, no. 3, pp. 443–455, 2014.

[12] S. Rayana and L. Akoglu, "Collective opinion spam detection: Bridging review networks and metadata," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 985–994.

[13] S. Kc and A. Mukherjee, "On the temporal dynamics of opinion spamming: Case studies on Yelp," in *Proceedings of the 25th International Conference on World Wide Web*, 2016, pp. 369–379.

[14] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What Yelp fake review filter might be doing?" in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, no. 1, 2013.

[15] Y. Zhu, H. Liu, Y. Du, and Z. Wu, "Ifspard: An information fusion-based framework for spam review detection," in *Proceedings of the Web Conference 2021*, 2021, pp. 507–517.

[16] S. Shehnepoor, M. Salehi, R. Farahbakhsh, and N. Crespi, "Netspam: A network-based spam detection framework for reviews in online social media," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1585–1595, 2017.

[17] Y. Yao, B. Viswanath, J. Cryan, H. Zheng, and B. Y. Zhao, "Automated crowdturfing attacks and defenses in online review systems," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1143–1158.

[18] Yelp, "Why do reviews change from "recommended" to "not currently recommended" and vice versa?" [Online; accessed 06-October-2021]. [Online]. Available: https://www.yelp-support.com/article/Why-do-reviews-change-from-recommended-to-not-currently-recommended-and-vice-versa

[19] S. Bakhshi, P. Kanuparthy, and E. Gilbert, "Demographics, weather and online reviews: A study of restaurant recommendations," in *Proceedings of the 23rd International Conference on World Wide Web*, 2014, pp. 443–454.

[20] J. Ye, S. Kumar, and L. Akoglu, "Temporal opinion spam detection by multivariate indicative signals," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 10, no. 1, 2016.

[21] C.-C. Wang, M.-Y. Day, C.-C. Chen, and J.-W. Liou, "Temporal and sentimental analysis of a real case of fake reviews in Taiwan," in *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2017, pp. 729–736.

[22] S. Nilizadeh, H. Aghakhani, E. Gustafson, C. Kruegel, and G. Vigna, "Think outside the dataset: Finding fraudulent reviews using cross-dataset analysis," in *The World Wide Web Conference*, 2019, pp. 3108–3115.

[23] I. Sutherland, Y. Sim, S. K. Lee, J. Byun, and K. Kiatkawsin, "Topic modeling of online accommodation reviews via latent dirichlet allocation," *Sustainability*, vol. 12, no. 5, p. 1821, 2020.

[24] M. H. van Velthoven, H. Atherton, and J. Powell, "A cross sectional survey of the UK public to understand use of online ratings and reviews of health services," *Patient Education and Counseling*, vol. 101, no. 9, pp. 1690–1696, 2018.

[25] M. Ott, C. Cardie, and J. Hancock, "Estimating the prevalence of deception in online review communities," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 201–210.

[26] E. Dwoskin and C. Timberg, "How merchants use Facebook to flood Amazon with fake reviews," *The Washington Post*, 2018. [Online]. Available: https://www.washingtonpost.com/business/economy/how-merchants-secretly-use-facebook-to-flood-amazon-with-fake-reviews/2018/04/23/5dad1e30-4392-11e8-8569-26fda6b404c7_story.html?

[27] C. Elliott, "This is why you should not trust online reviews," *Forbes*, 2018. [Online]. Available: https://www.forbes.com/sites/christopherelliott/2018/11/21/why-you-should-not-trust-online-reviews/\#2f36c6d52218
[28] Yelp, "Does Yelp recommend every review?" [Online; accessed 18-October-2020]. [Online]. Available: https://www.yelp-support.com/article/Does-Yelp-recommend-every-review
[29] D. Martens and W. Maalej, "Towards understanding and detecting fake reviews in app stores," *Empirical Software Engineering*, vol. 24, no. 6, pp. 3316–3355, 2019.
[30] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. Subrahmanian, "Fraudulent user prediction in rating platforms," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018, pp. 333–341.
[31] C. G. Harris, "Detecting deceptive opinion spam using human computation," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.
[32] D. I. Adelani, H. Mai, F. Fang, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection," in *International Conference on Advanced Information Networking and Applications*. Springer, 2020, pp. 1341–1354.
[33] M. Juuti, B. Sun, T. Mori, and N. Asokan, "Stay on-topic: Generating context-specific fake restaurant reviews," in *European Symposium on Research in Computer Security*. Springer, 2018, pp. 132–151.
[34] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 309–319.
[35] C.-C. Wang, M.-Y. Day, and Y.-R. Lin, "A real case analytics on social network of opinion spammers," in *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*. IEEE, 2016, pp. 623–630.
[36] Yelp, "Recommendation software," [Online; accessed 03-October-2021]. [Online]. Available: https://trust.yelp.com/recommendation-software/
[37] ——, "Do reviews that aren't currently recommended impact the business's star rating?" [Online; accessed 8-February-2021]. [Online]. Available: https://www.yelp-support.com/article/Do-reviews-that-arent-currently-recommended-impact-the-businesss-star-rating
[38] ——, "Yelp open dataset," [Online; accessed 07-October-2021]. [Online]. Available: https://www.yelp.com/dataset
[39] American Community Survey, "Household income in the past 12 months (in 2019 inflation-adjusted dollars)," 2019, accessed: 2021-10-13. [Online]. Available: https://data.census.gov/cedsci/table?tid=ACSDT5Y2019.B19001
[40] U. S. C. Bureau, "ZIP Code Tabulation Areas (ZCTAs)," August 2020, [Online; accessed 05-October-2021]. [Online]. Available: https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html
[41] American Community Survey, "Total population," 2019, accessed: 2021-10-13. [Online]. Available: https://data.census.gov/cedsci/table?tid=ACSDT5Y2019.B01003
[42] Yelp, "Yelp Fusion," [Online; accessed 18-January-2021]. [Online]. Available: https://www.yelp.com/fusion
[43] pyppeteer, "miyakogi/pyppeteer: Headless chrome/chromium automation library (unofficial port of puppeteer)," May 2019, [Online; accessed 15. Apr. 2020]. [Online]. Available: https://github.com/miyakogi/pyppeteer
[44] D. Altig, S. Baker, J. M. Barrero, N. Bloom, P. Bunn, S. Chen, S. J. Davis, J. Leather, B. Meyer, E. Mihaylov *et al.*, "Economic uncertainty before and during the COVID-19 pandemic," *Journal of Public Economics*, vol. 191, p. 104274, 2020.
[45] P. Deb, D. Furceri, J. D. Ostry, and N. Tawk, "The economic effects of Covid-19 containment measures," 2020.
[46] Yelp, "Yelp updates recommendation software to better target and mitigate content from online review exchange groups," Feb 2021, [Online; accessed 15-October-2021]. [Online]. Available: https://blog.yelp.com/news/yelp-updates-recommendation-software-to-better-target-and-mitigate-content-from-online-review-exchange-groups/

APPENDIX

A. Post-processing and organization

We took additional steps to clean up and organize our data.

Deduplication. Our data has some duplicates. Some of these may be real; for example, if the author accidentally submitted the review twice. However, because our crawls were not instantaneous, review order sometimes shifted during crawling, occasionally leading to double collection of the same review. In either case, such reviews may affect the accuracy of the analysis. Thus we removed these reviews. To remove duplicate reviews, we removed reviews where all fields (e.g. text, author, date) are identical, retaining one copy.

In our CHI-3 crawl, approximately 85,000 reviews appeared under both Recommended and Not Recommended, and appeared under Recommended for the adjacent crawls (CHI-2/CHI-4). This coincides with a major update to the Yelp recommendation software [46]. This event boosts the number of double reclassifications approximately 80-fold if we treat these reviews as Not Recommended, since they are Recommended in CHI-3, Not Recommended in CHI-4, and Recommended in CHI-5. To address this anomaly, we keep the Recommended version of each review and discard the Not Recommended version.

Matching reviews. We do not have a unique identifier for reviews, so we rely on heuristics to identify instances of the same review across crawls. To determine if two reviews match, we find all reviews with the same text. If two such reviews appear in the same crawl, we discard all reviews with that text, because we cannot disambiguate them (0.04% of reviews for CHI and 0.04% for UDIS). Otherwise, we assume the reviews with that text are the same review.

Determining authorship. Unlike prior work [12], [14], we were unable to find a universal identifier for authors. Instead, we found two sets of author identifiers: one for Recommended reviews and one for Not Recommended and Removed reviews. This may be due to site design changes on Yelp. We considered matching authors based on metadata but observed too many false positives to consider this approach reliable. However, for authors with at least one reclassified review, the combination of both identifiers serves as a universal identifier. Thus we focus our investigation of authorship on authors with at least one reclassified review.

B. Collected data

TABLE IV
DATA AND METADATA COLLECTED.

Field	Description
Reviews	
Content	Text of the review
Author ID	Reviewer identifier (differs for Recommended / Not Recommended reviews)
Date	Date of posting
Rating	Review rating
Business ID	Identifier for the business the review was posted to
Author data	Name and other public account information
Recommended	Whether the review is Recommended
Businesses	
Business ID	Identifier for the business
Amenities	Listed amenities

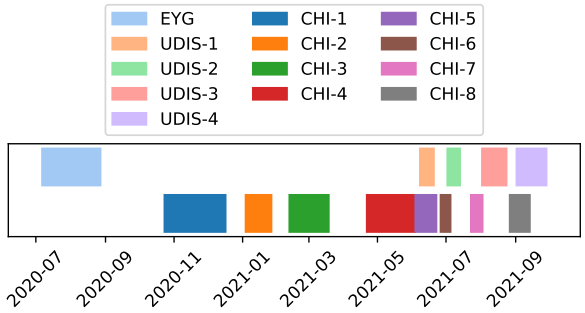

Fig. 6. The timeline for each crawl. Each box indicates the first and last operation for each crawl.

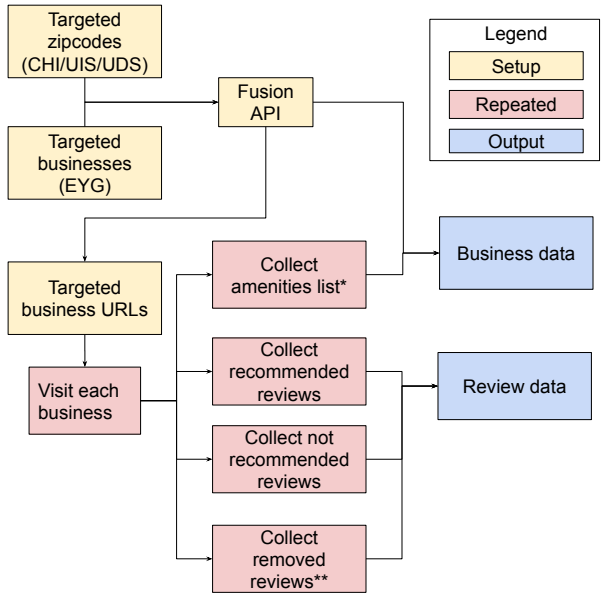

Fig. 7. The data collection process. Yellow indicates setup steps that are completed once. Red indicates steps that are completed for each crawl. Blue indicates outputs.
* Amenities were collected only for CHI 8 and UDIS 4.
** Removed reviews were collected for CHI 7-8 and UDIS 3-4.