Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Can Social Networks Substitute for Reputation Systems? Evidence from a Large Scale Field Experiment

David Holtz[1, 2], **P. Alex Dow**[3], **Sinan Aral**[4, 2]
[1]Columbia Business School
[2]MIT Initiative on the Digital Economy
[3]Microsoft Research
[4]MIT Sloan School of Management
david.holtz@columbia.edu, alex.dow@microsoft.com, sinan@mit.edu

Abstract

Reputation systems are considered by many to be foundational to online peer-to-peer marketplaces. Both economic theory and previous empirical work suggest that centralized reputation should increase trust, increase trustworthiness, and foster marketplace growth. However, prior work also suggests that embedding peer-to-peer commerce in social settings can create trust and obviate the need for centralized reputation. Thus, it is unclear whether we should expect centralized reputation to matter in a setting where social network features are also generating trust. In this paper, we present results from what we believe is the first ever field experiment to randomize the introduction of a reputation system into a massive online marketplace. The setting for our study is an online peer-to-peer marketplace embedded inside of a large social network. We find that on average, the introduction of ratings did not have an effect on the amount of marketplace activity, the behavior of marketplace users, or the structure of the marketplace in the six months following its introduction. Our results suggest that centralized reputation may be unnecessary when online marketplaces are embedded in social networks. Given the number of previously documented biases that can arise in centralized reputation systems, the use of social network data may provide an attractive alternative for online marketplace designers.

Introduction

Digital peer-to-peer marketplaces, such as Airbnb, eBay, and Upwork are an increasingly common venue for people to exchange goods, experiences, and labor. Participating in these markets confers benefits to both shoppers and sellers. For instance, sellers are able to sell their wares without incurring the large fixed costs associated with starting a dedicated eCommerce website or opening a brick-and-mortar store, while shoppers are able to discover a long tail of sellers they may have otherwise never encountered. Early digital peer-to-peer marketplaces were met with skepticism, since the success of an online marketplace is predicated on successfully establishing trust between shoppers and sellers. Adverse selection and moral hazard concerns (Akerlof 1978) are even more pronounced in online markets than in offline markets, because in many digital peer-to-peer marketplaces people are interacting with strangers and/or transacting under pseudonyms. Furthermore, the relative anonymity of early online marketplaces reduced the potential reputational consequences of opportunistic behavior.

Digital marketplaces are well-suited to collect detailed information on shopper and seller behavior, and costlessly transmit that information to future shoppers and sellers (Resnick et al. 2006). Given this fact, most online marketplace designers have mitigated trust concerns through the introduction of centralized reputation systems.[1] Theoretical models suggest that the introduction of centralized reputation should facilitate transactions and lead to marketplace growth (Kreps 1996; Tadelis 2016), and previous lab experiments (Bohnet and Huck 2004; Bolton, Katok, and Ockenfels 2004; Bolton, Greiner, and Ockenfels 2013) and observational studies (Cai et al. 2014) have shown that introducing a reputation system can increase trust, increase trustworthiness, and affect market dynamics. However, prior work (Holtz, MacLean, and Aral 2017) also suggests that embedding peer-to-peer commerce in a social network may create trust, and thus substitute for or obviate the need for centralized reputation systems. Therefore, it is not clear whether we should expect centralized reputation to have an impact in an online marketplace where social network features are also generating trust.

In this paper, we conduct the first ever field experiment to randomize the introduction of a reputation system into a massive online marketplace, and proceed to study the impact of reputation systems in settings where social network information may foster trust, thus reducing the need for centralized reputation. The setting for our study is Rialto (name disguised), which is embedded inside of a large social network that we call SocialNet (name disguised). Rialto is a section of SocialNet's website and apps that enables SocialNet users to buy and sell goods from others in their local community. The goods available on Rialto span a wide range of categories, including clothing, furniture, electronics, and vehicles. In most of the world, Rialto did not have a centralized reputation system prior to December 2017.[2] Beginning on December 5, 2017, SocialNet conducted a randomized,

[1]See Bar-Isaac, Tadelis et al. (2008) for a survey of the early theoretical work on reputation systems.

[2]A beta version of the reputation system that serves as our experimental treatment was available in Australia and Washington State prior to December 2017.

state-by-state rollout of a centralized, bilateral reputation system for Rialto. Using a matched pairs design, twenty-five states were assigned to the treatment condition, whereas twenty-four states were assigned to the control condition. The reputation system was introduced to treatment states on one of three dates between December 5, 2017 and February 21, 2018, whereas the reputation system was introduced to control states on February 22, 2018. The exogenous variation in the timing with which centralized reputation is introduced to different states allows us to estimate the causal impact of Rialto's reputation system.[3]

We find that, contrary to evidence from marketplaces without social network information, in our setting, centralized reputation had no overall effect on Rialto after being live for six months. We analyze the results of our experiment using a difference-in-differences estimator, and find that centralized reputation did not have a statistically significant effect on the number of items posted to Rialto, the number of users visiting Rialto, the number of conversations initiated on Rialto, the number of double interacts on Rialto, or the amount of reported bad behavior on Rialto. We also find that the introduction of ratings did not have an effect on market concentration, and did not affect the price distribution of items posted to Rialto. These results suggest that embedding peer-to-peer commerce inside of social networks can substitute for the trust that centralized reputation systems create in the absence of explicit social connections. Given the fact that reputation systems are susceptible to a number of biases (Muchnik, Aral, and Taylor 2013; Fradkin, Grewal, and Holtz 2021; Nosko and Tadelis 2015; Filippas, Horton, and Golden 2022), the possibility that social network information can substitute for centralized reputation has serious implications for online marketplace designers.

In addition to the social embeddedness of Rialto, there are additional factors that may explain our results. First, the initial design of Rialto's reputation system may not have effectively solicited feedback, or may not have displayed existing feedback prominently enough. Second, it is possible that introducing centralized reputation to a digital marketplace has a long-run overall effect, but no short-run overall effect. While our treatment effect point estimates do provide suggestive, non-statistically significant evidence that the impact of centralized reputation is positive and increases over time, we do not have the statistical power to make strong claims on this topic. Given these alternative explanations, we believe that further research is needed to definitively determine under what circumstances social networks can substitute for or obviate the need for centralized reputation systems.

Related Literature

Much of the early work on reputation systems established the theoretical motivation for reputation systems, and established that centralized reputation can help to alleviate market failures stemming from adverse selection and moral hazard. We refer the reader to Bar-Isaac, Tadelis et al. (2008) and Dellarocas (2003) for a review of this work. Since online marketplaces, such as eBay, began incorporating reputation systems into their design in the 2000s, a number of observational studies (Livingston 2005; Jin and Kato 2006; Cabral and Hortacsu 2010; Filippas, Horton, and Golden 2022) and field experiments (Resnick et al. 2006; Fradkin, Grewal, and Holtz 2021; Garg and Johari 2021; Klein, Lambertz, and Stahl 2016) have focused on reputation in online marketplaces.

In the reputation systems literature, this study relates most closely to two papers examining market-level reputation effects.[4] Cai et al. (2014) study the introduction of centralized reputation on Eachnet.com, finding that reputation systems reduce differences between high and low reputability sellers while enabling high-reputation sellers to expand into new markets. Bolton, Katok, and Ockenfels (2004) conduct laboratory experiments comparing strangers markets (no feedback), partners markets (repeated interaction), and feedback markets (reputation system). They find that while feedback improves efficiency and trust relative to strangers markets, direct repeated interaction performs best, suggesting that social relationships may be superior to formal reputation systems.

This paper also contributes to research on commerce embedded in social structure. Granovetter (1985) argues that embedding economic activity in social structure encourages trust between transaction partners and reduces the probability of malfeasance, a perspective reinforced by Coleman (1988), who shows that dense social networks facilitate cooperation through monitoring and social sanctions. Holtz, MacLean, and Aral (2017) conducted an observational study of peer-to-peer commerce in Facebook Buy & Sell Groups, finding that increased network density and seller network centrality both lead to increased two-sided demand between shoppers and sellers. An important difference between our marketplace and other online peer-to-peer platforms is its embeddedness within a pre-existing social network. This social embeddedness may obviate the need for formal reputation systems, as shoppers and sellers already have social information to assess potential transaction partners' trustworthiness.[5] Additionally, users transact with their real social network accounts rather than anonymous profiles, making reputation more persistent since users cannot easily delete accounts and create new ones (Friedman and Resnick 2001).

Like Bolton, Katok, and Ockenfels (2004) and Cai et al. (2014), this paper measures the impact of centralized reputation on a marketplace. However, our study is the first

[3]This research was deemed exempt by an Institutional Review Board as it involves analysis of anonymized behavioral data from a staggered rollout conducted by Rialto for business purposes. Data analysis was conducted on company systems with deidentified user data and appropriate access controls.

[4]There is also a large literature on seller-level impacts (Cabral and Hortacsu 2010; Resnick et al. 2006).

[5]An internal experiment conducted by SocialNet in January 2018 provided additional evidence supporting this hypothesis. In the experiment, user commerce profiles in Rialto included either social information about the seller, non-social information about the seller, only basic information about the seller, or a combination of the three. Relative to the other conditions, shoppers who saw social information about the seller were more likely to initiate a Rialto conversation with the seller.

Metric	Min	Median	Mean	Max	Std. Dev.	N_{obs}
Rialto visitors	143,448	1,378,008	1,724,000	8,950,424	1,777,256	735
Items posted	1,184	28,576	39,636	246,484	41,548	735
Product detail page views	1,188,752	13,973,668	18,943,904	115,375,904	20,047,628	735
Conversations initiated	8,356	143,320	190,468	1,252,684	207,140	735
Double interacts	3,972	66,060	87,964	524,656	90,536	735
Items marked as sold	708	14,608	19,780	112,696	19,344	735
Reported sellers	4	304	424	2,620	448	735
Reported shoppers	0	80	116	660	120	735
Median Price (USD)	20	40	43	99	10	735
50%/10%	4	6	7	13	1	735
90%/10%	44	101	116	510	47	735
90%/50%	10	16	18	60	6	735
Herfindahl index	3×10^{-5}	0.00022	0.00041	0.00318	0.00051	392
Normalized Herfindahl index	3×10^{-5}	0.00019	0.00036	0.00293	0.00045	392

Table 1: Descriptive statistics for Rialto in the U.S. from November 20th, 2017 to December 4th, 2017. Each observation is a U.S. state-day pair. All states except for Washington are included in the sample. Market concentration metrics have fewer pre-treatment observations because of smoothing. Note that values for the counting statistics have been scaled by an unreported constant to maintain the confidentiality of financially sensitive Rialto data while preserving relative proportions and statistical relationships.

to randomize the introduction of a reputation system into a large online marketplace, and also the first to examine the impact of centralized reputation when a peer-to-peer market is embedded in a social network. Given previous work suggesting that social network structure can foster trust (Holtz, MacLean, and Aral 2017), it is unclear whether a centralized reputation system will still have an effect.

Setting and Descriptive Statistics

Rialto is a section of SocialNet's website and apps that describes itself as a destination for SocialNet users to buy and sell goods from others in their local community. Rialto allows users to buy and/or sell a wide range of items, including clothing, furniture, electronics, and vehicles. Every interaction on Rialto involves two parties: the "Shopper," who is interested in purchasing a given item, and the "Seller," who is attempting to sell that item. Shoppers could browse items for sale within a defined radius, filter by category and price, and view product details including photos, descriptions, and seller information (e.g., mutual friends, responsiveness). If interested, shoppers could message sellers to coordinate transaction details through custom Rialto conversation threads. Sellers posted items by providing a title, description, price, category, location, and photos.

After completing a transaction (often offline), sellers were encouraged to mark an item as sold. Once an item was marked as sold, it ceased to appear in search for Rialto shoppers. SocialNet does not process payments for transactions conducted on Rialto, so in cases where a seller does mark an item as sold, it is generally difficult to determine which shopper the item was sold to. An item being marked as sold is also an unreliable indicator of whether or not that item was actually sold on Rialto; some sellers who do sell an item on Rialto neglect to mark it as sold (for instance, they may simply delete the item), whereas sellers who mark an item as sold may have failed to sell the item or sold the item elsewhere. Prior to and after the launch of ratings, both shoppers and sellers are able to report their counterparty to SocialNet if they feel their counterparty has acted inappropriately.

Categorizing Rialto items as particular goods in an automated fashion is difficult for two reasons. First, there are no guarantees that the name, description, or photos provided by a seller describe the item being sold well. For instance, if a seller is selling an iPhone, it may be difficult to determine this from the provided information. Even if it is clear the item being sold is an iPhone, it may not be clear what specific version is being sold. Second, many of the items for sale are rare, one-of-a-kind items that cannot be matched to items for sale elsewhere in the physical or digital world. Although identifying particular items for sale on Rialto is difficult, sellers were required to assign items for sale to one of thirty-two categories. Example item categories include "arts and crafts," "tools," "electronics and computers," "cars, trucks and motorcycles," and "property for sale."[6]

It is important to note that, at the time of our experiment, most Rialto sellers in the U.S. were "casual sellers." Among sellers who posted an item to Rialto between November 20, 2017 and May 1, 2018, 89% posted 10 or fewer items. Because most sellers are casual, sellers that *do* post multiple items are unlikely to be selling multiple copies of the same item. Taking this information into account, most interactions on Rialto are one-shot, and "reputation" on Rialto (broadly defined) is likely to quantify the likelihood that a shopper or seller will execute a transaction smoothly and not behave opportunistically. For sellers in particular, reputation on Rialto is unlikely to quantify the quality of the good or service being sold.

[6]It is worth noting that item categories tend to include accessories and paraphernalia. For example, many items for sale in the "mobile phones" category are phone chargers, phone cases, and other mobile phone-related items.

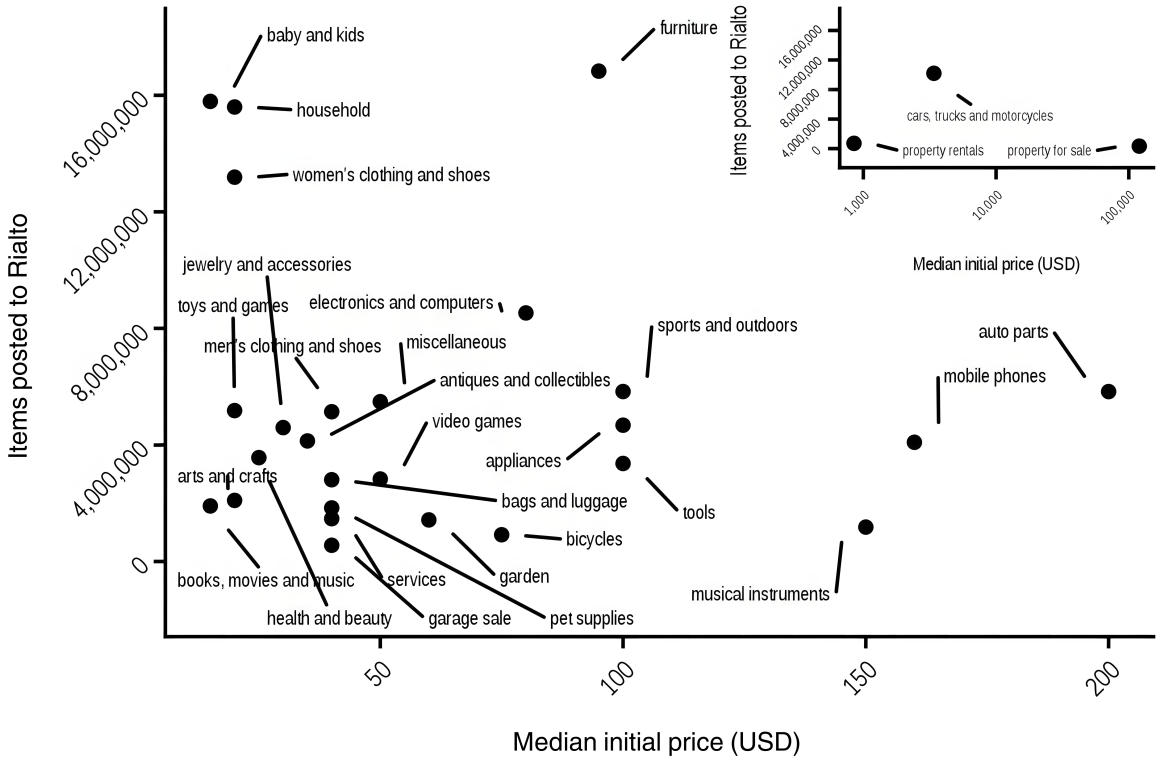

Figure 1: The number of items posted and median initial price in each Rialto category between November 20, 2017 and December 4, 2017. The inset plot's x-axis is on a log scale.

Descriptive Statistics

Table 1 shows descriptive statistics for Rialto in 49 U.S. states (all except Washington) from November 20, 2017 to December 4, 2017, i.e., the fifteen days immediately preceding the staggered rollout of ratings to Rialto in the U.S. Each observation is a state-day pair, producing a dataset of 735 observations. We report summary statistics for the number of Rialto visitors, items posted, product detail page views, conversations initiated, double interacts,[7] items marked as sold, reported sellers, and reported shoppers per state-day over that period.[8] There are many more conversations initiated per state-day than there are items posted, indicating that on average sellers receive messages from multiple shoppers after posting an item. On average, 44% of those conversations become double interacts, indicating that sellers are responsive and suggesting that shoppers have genuine interest in transacting. Notably, the number of reported shoppers and sellers is very small relative to the amount of activity occurring on Rialto. This could indicate that most shoppers and sellers do not report inappropriate behavior when it occurs, or that there is little inappropriate behavior on Rialto for shoppers and sellers to report.

Figure 1 shows the number of items posted and the median price for each Rialto item category from November 20, 2017 to December 4, 2017.[9] The most common item categories on Rialto are "furniture," "baby and kids," "household," "woman's clothing and shoes," and "cars, trucks and motorcycles." The most expensive categories are "property for sale," "cars, trucks and motorcycles," and "property rentals," whereas the least expensive categories are "arts, movies and music," "women's clothing and shoes," and "baby and kids."

[7]The double interact is a pattern of interaction first introduced by (Weick 1979). The double interact can be described as a pattern of messaging behavior in which agent A messages agent B, agent B responds to agent A, and agent A responds to agent B's response. The double interacts has previously been used as a proxy for "transactions" in studies of online dating markets (Bapna et al. 2016) and two-sided interest in transacting in studies of Facebook Buy & Sell Groups (Holtz, MacLean, and Aral 2017).

[8]Values for the counting statistics have been scaled by an unreported constant to maintain the confidentiality of financially sensitive Rialto data while preserving relative proportions and statistical relationships. The same unreported constant is used consistently throughout the entire paper.

[9]We omit three categories ("housing," "electronic cases," and "power adapters") from this chart and our subsequent analysis, since the number of items in these categories were orders of magnitude smaller than the other 29 categories. The largest of these three categories ("housing") had 10 posts over the time period in question, making it orders of magnitude smaller than the next smallest category ("property for sale").

	Dependent variable:		
	log(visitors)	log(items posted)	log(product detail page views)
	(1)	(2)	(3)
Ratings	−0.003	−0.003	−0.006
	(−0.011, 0.005)	(−0.029, 0.023)	(−0.022, 0.010)
Observations	7,987	7,987	7,987
State-level F.E.	Yes	Yes	Yes
Time F.E.	Yes	Yes	Yes
R^2	0.001	0.0001	0.001

*p<0.1; **p<0.05; ***p<0.01

Note: Constant term omitted to maintain the confidentiality of financially sensitive Rialto data.

Table 2: The overall effect of ratings on market activity-related metrics

Ratings Rollout Experiment

On December 5, 2017, SocialNet began the staggered rollout of a ratings system to Rialto across 49 U.S. states. In this section, we describe the design of that ratings system, as well as the experiment design that dictated its staggered rollout.

Ratings Implementation

The ratings system rolled out to Rialto beginning on December 5, 2017 allowed shoppers and sellers to rate each other. Rialto ratings were developed as a 'mobile-first' feature. As a result, ratings were not visible on Desktop computers, and both shoppers and sellers were only able to enter the rating flow on mobile. An important distinction between the Rialto ratings system and other bilateral reputation systems is that shoppers and sellers did not need to have completed a transaction with one another in order to rate each other. Sellers were able to rate any shopper who messaged them about an item in Rialto, whereas shoppers could rate any seller who responded to their initial inquiry.

There were three different ways for shoppers and sellers to enter the rating flow. First, shoppers and sellers had the option to rate their counterparty by clicking a call to action in the top right corner of the conversation thread with their counterparty. Second, if a seller marked an item as sold, they were able to choose which shopper they had sold the item to, after which they were asked if they would like to rate that shopper. Finally, if either the shopper or seller received a rating, a notification appeared at the top of their Rialto feed. This notification informed them that they had been rated by their counterparty, and asked if they would like to rate them back.

After choosing to rate their counterparty, sellers were solicited for binary feedback. They were asked whether they would recommend their counterparty as a buyer, and were able to choose one of two icons: a frowning face (indicating no), or a smiling face (indicating yes). In the event that the seller chose the frowning face, they were asked how the shopper can improve. They were able to choose as many as six of the following choices: "Punctuality," "Friendliness," "Negotiation," "Response Time," "Reliability," and "Payment Speed." The flow for shoppers rating sellers was identical, with the exception that shoppers were asked whether they would recommend their counterparty as a seller. Rating information was displayed on a user's commerce profile, which was a separate profile from their primary SocialNet profile. The user commerce profile was accessible by clicking on a seller's name in the "Seller Information" section of the product details page, and was the only place where ratings were displayed. When viewing their own profile, Rialto users were able to see the cumulative count of positive and negative ratings they had received as both a shopper and a seller, as well as the number of times each area for improvement had been suggested by those who rated them negatively. Only ratings received as a seller were visible to others, and users had the options of hiding their seller ratings so they did not appear on their commerce profile.

Experiment Design

Pre-treatment Rialto data was collected from June 9, 2017 to August 5, 2017. Each unique seller that posted an item to Rialto during this time was assigned to a state based on the bucketed latitude and longitude of the first item they posted. Each item was then assigned its seller's U.S. state. We aggregated data at the state level, and for each state calculated the number of items posted, product detail page views, conversations initiated, double interacts, items marked as sold, reported shoppers, and reported sellers in Rialto from June 9, 2017 to July 6, 2017. We also calculated the number of items marked as sold in non-Rialto, commerce-related groups on SocialNet in each state between June 9, 2017 and July 6, 2017. Finally, we calculated the number of unique sellers active in each state between June 9, 2017 and August 5, 2018.

States were then arranged into pairs using a multivariate blocking procedure (Moore 2012). First, each of the aforementioned metrics was centered and scaled. We then calculated the Mahalanobis distance between each pair of states, and arranged states into pairs using an optimal-greedy algorithm that repeatedly finds the two remaining states with the smallest Mahalanobis distance. Within each pair, one state was randomly designated a treatment state, and the other

	Dependent variable:		
	log(conversations initiated)	log(double interacts)	log(items marked as sold)
	(1)	(2)	(3)
Ratings	−0.009	−0.006	−0.023*
	(−0.028, 0.011)	(−0.024, 0.012)	(−0.045, 0.0003)
Observations	7,987	7,987	7,987
State-level F.E.	Yes	Yes	Yes
Time F.E.	Yes	Yes	Yes
R^2	0.001	0.001	0.004

*p<0.1; **p<0.05; ***p<0.01

Note: Constant term omitted to maintain the confidentiality of financially sensitive Rialto data.

Table 3: The overall effect of ratings on Rialto transaction-related metrics

state was designated a control state.[10]

States assigned to the treatment condition were included in a staggered rollout of Rialto ratings that began on December 5, 2017 and concluded on February 21, 2018. This led to the following rollout schedule for Rialto ratings. On December 5, 2017, ratings were introduced in Arkansas, Florida, and Idaho. On January 9, 2018, ratings were introduced in Colorado, Mississippi, and West Virginia. On January 25, 2018, ratings were introduced to Delaware, Illinois, Kansas, Kentucky, Maine, Maryland, Massachusetts, Minnesota, Montana, New Hampshire, New Jersey, New Mexico, New York, North Dakota, Oklahoma, Pennsylvania, Rhode Island, Tennessee, and Texas. On February 22nd, 2018, ratings were introduced to the full set of control states, i.e., all remaining U.S. states except Washington.[11] Unfortunately, the timing with which different states in the treatment gained access to ratings was not randomized, although members of SocialNet's product team attempted to choose the rollout order in such a way that it was *essentially* random subject to business constraints.[12]

Figure 3 (in the Appendix) shows the average number of items posted per state-day from November 20, 2017 to May 1, 2018 in both treatment and control groups. We inspected similar parallel trends plots for metrics such as visits, conversations initiated, double interacts, and reported users, although we omit them due to space constraints (additional figures available upon request). The treatment and control exhibit parallel trends for each metric, indicating that our randomization procedure worked well and that a difference-in-differences-style analysis is appropriate. The average values of most metrics are extremely similar in the treatment and control arms prior to December 5, providing further evidence that our randomization procedure was effective.[13]

Figure 4 (in the Appendix) displays aggregate ratings-related descriptive statistics for the 49 states in our sample over time. The blue, green, red, and magenta vertical lines correspond to each of the four dates that ratings were introduced to new states. As expected, the number of shopper and seller ratings increase immediately after each of the dates in the staggered rollout plan, and again after ratings are released in the control states. Consistently, over 80% of shopper ratings and over 85% of seller ratings are positive. This is consistent with evidence from other bilateral review systems (Zervas, Proserpio, and Byers 2021). Finally, it is worth noting that rate at which both shoppers and sellers rate their counterparties is extremely low ($< 3\%$).

Experiment Results

In this section, we present the results of the ratings rollout experiment. Our analysis dataset consists of aggregate statistics for each state-day from November 20, 2017 to May 1, 2018.

Activity Metrics

We first present the effect of ratings on outcomes that describe activity levels on Rialto. We group these outcomes into three categories: outcomes related to market activity (Rialto visitors, items posted to Rialto, and product detail page views), outcomes related to transacting (conversations initiated, double interacts, and items marked as sold), and outcomes related to reporting (reported shoppers and reported sellers).

We use a difference-in-differences estimator to estimate the effect of introducing ratings on each of these market outcomes. More specifically, we estimate regressions of the form:

$$Y_{st} = \alpha_s + \gamma_t + \tau \mathbf{1}\left(\text{ratings}_{st} > 0\right) + \epsilon_{st}, \qquad (1)$$

[10]Table 7 (in the appendix) shows the twenty-four state pairs, the Mahalanobis distance between the two states in each pair, and the treatment/control assignments within each pair. The number of states we provided to the algorithm was odd. As a result, Texas went unmatched.

[11]Prior to February 22, 2018, 5% of users in treatment states were part of a "holdout" group that could not see or leave ratings.

[12]The product team wanted to ensure that large U.S. states, such as Texas or California, were not among the first states to receive access to ratings.

[13]The exception is average Rialto visitors per state-day. While the average number of visitors per day in the treatment and control arm exhibit parallel trends, the average number of visitors differ.

	Dependent variable:	
	log(reported shoppers + 1)	log(reported sellers + 1)
	(1)	(2)
Ratings	-0.064^{*}	0.034
	$(-0.129, 0.001)$	$(-0.009, 0.076)$
Observations	7,987	7,987
State-level F.E.	Yes	Yes
Time F.E.	Yes	Yes
R^2	0.001	0.001

*p<0.1; **p<0.05; ***p<0.01

Note: Constant term omitted to maintain the confidentiality of financially sensitive Rialto data.

Table 4: The overall effect of ratings on Rialto reporting-related metrics

where Y_{st} is the dependent variable in state s at time t, ratings_{st} is the number of days that ratings have been live in a given state, and the indicator variable $\mathbf{1}\left(\text{ratings}_{st} > 0\right)$ is 1 when ratings are live in a given state, and 0 when they are not. α_s is a state-level fixed effect, γ_t is a day-level fixed effect, and ϵ_{st} is the error term. The two-way fixed effects specification controls for time-invariant characteristics of Rialto in different states, as well as common time trends and daily fluctuations in our outcome metrics. Standard errors are clustered at the state level to account for any serial correlation in outcomes (Bertrand, Duflo, and Mullainathan 2004), and are calculated using the wild cluster bootstrap with $n = 1,000$ replications (Cameron, Gelbach, and Miller 2008).[14]

Table 2 shows the estimated effect of ratings on log-transformed market activity-related outcomes, and Table 3 shows the estimated effect of ratings on log-transformed transaction-related outcomes. The introduction of ratings does not have a statistically significant impact on log(visitors) ($p = 0.44$), log(items posted) ($p = 0.82$), log(product detail page views) ($p = 0.46$), log(conversations initiated) ($p = 0.39$), log(double interacts) ($p = 0.51$), or log(items marked as sold) ($p = 0.06$). These null results are precisely estimated.[15]

Table 4 shows the estimated effect of ratings on log(reported shoppers + 1) and log(reported sellers + 1). The introduction of ratings does not have a statistically significant impact on log(reported shoppers + 1) ($p = 0.053$) or log(reported sellers + 1) ($p = 0.113$). These null results are less precisely estimated than our null results for activity metrics in Table 2 because there are so few shopper and seller reports on Rialto, limiting our statistical power to detect treatment effects.[16]

Market Structure Metrics

In addition to measuring the effect of introducing ratings on outcomes that quantify Rialto's market activity levels, we measure the effect of introducing ratings on the market structure of Rialto. More specifically, we test for whether the introduction of ratings had a causal effect on the price distribution of items posted to Rialto or the level of market concentration found on Rialto.

We first consider changes to the price distribution of goods on Rialto. Although sellers on Rialto are selling heterogeneous goods across multiple item categories, testing for changes in the distribution of prices should provide a rough sense of whether ratings change the type of items that sellers post to Rialto. For our price-related analyses, we omit items with prices set to zero, and items with prices above $1M USD.[17] We estimate the effect of ratings on four price distribution-related outcomes, each measured at the state-day level: the median price of items posted to Rialto, the ratio of the 90th percentile price to the 10th percentile price,

[14]Recent work has shown that two-way fixed effects estimators in staggered adoption designs can assign negative weights to some group-time treatment effects (Goodman-Bacon 2021; de Chaisemartin and D'Haultfœuille 2020; Callaway and Sant'Anna 2021). Our randomization inference analysis (Table 8), which exploits the random assignment of treatment timing across states rather than relying on TWFE, independently confirms our null findings.

[15]We convert our estimated coefficients and 95% confidence intervals to percentage changes with the transformation $\tau_{pct} = exp(\tau) - 1$. The 95% confidence intervals in percentage terms are as follows: log(visitors): (-1.09%, +0.50%), log(items posted): (-2.86%, +2.32%), log(product detail page views): (-2.17%, +1.01%), log(conversations initiated): (-2.76%, +1.11%), log(double interacts): (-2.37%, +1.21%), log(items marked as sold): (-4.40%, +0.03%), log(reported shoppers + 1): (-12.10%, +0.10%), log(reported sellers + 1): (-0.90%, +7.90%)

[16]To verify the robustness of market activity results, we re-estimated Equation 1 both with inverse hyperbolic sine transformed outcomes (Burbidge, Magee, and Robb 1988; MacKinnon and Magee 1990) and as a quasi-Poisson regression model (Silva and Tenreyro 2006; Wooldridge 2006). Our findings are robust to both of these modifications. A more detailed description of these analyses can be found in the Appendix, and results are available upon request.

[17]Sellers might list items with the price set to zero if they are giving something away for free, or listing multiple different items with one post. Prices above $1M are often typos, gibberish, or sellers using an item's price to communicate non-price information to shoppers.

	Dependent variable:			
	Median Price	$\frac{90\%}{10\%}$	$\frac{90\%}{50\%}$	$\frac{50\%}{10\%}$
	(1)	(2)	(3)	(4)
Ratings	0.238	3.489	0.317	0.035
	(0.448)	(3.347)	(0.483)	(0.055)
Observations	7,987	7,987	7,987	7,987
State-level F.E.	Yes	Yes	Yes	Yes
Time F.E.	Yes	Yes	Yes	Yes
R^2	0.793	0.691	0.711	0.372

*p<0.1; **p<0.05; ***p<0.01

Note: Constant term omitted to maintain the confidentiality of financially sensitive Rialto data.

Table 5: Price (dispersion) regression results

the ratio of the 90th percentile price to the median price, and the ratio of the median price to the 10th percentile price.

For each of these metrics, we estimate Equation 1, with standard errors clustered at the state-level. Table 5 shows our results. We find that the introduction of ratings did not have a statistically significant effect on the median price of items ($p = 0.60$), the ratio of 90th percentile price to median price ($p = 0.51$), the ratio of median price to 10th percentile price ($p = 0.53$), or the ratio of 90th percentile price to 10th percentile price ($p = 0.30$).

We next estimate the effect of introducing ratings on market concentration. In conventional industry settings, market concentration is typically measured using Herfindahl indices, which quantify how concentrated seller activity is—higher values indicate that a few sellers dominate posting, while lower values indicate more equal distribution across many sellers. The standard Herfindahl index ranges from 1/N (least concentrated) to 1 (most concentrated), while the normalized version adjusts for the number of active sellers and ranges from 0 to 1, making comparisons across different market sizes more meaningful. However, these standard metrics are less suitable for our marketplace context, where sellers post sporadically rather than maintaining persistent presence over time. To account for this sporadic posting behavior, we compute smoothed versions of both the standard and normalized Herfindahl indices that incorporate exponentially decayed past and future market participation within a seven-day window (see the Appendix for mathematical definitions).

We estimate the effect of ratings on both the smoothed Herfindahl index and the smoothed normalized Herfindahl index using three approaches: a linear probability model, logistic regression, and beta regression (Ferrari and Cribari-Neto 2004). Table 6 shows the estimated effect of ratings on both outcomes using all three approaches. We are unable to reject the null hypothesis of no ratings effect for either the smoothed Herfindal index (OLS: $p = 0.72$, logistic: $p = 0.86$, beta: $p = 0.80$) or the smoothed normalized Herfindahl index (OLS: $p = 0.73$, logistic: $p = 0.87$, beta: $p = 0.80$). We also re-estimate the effect using ratings with smoothing windows of three and five days, and find that our results are qualitatively and quantitatively similar (detailed results available upon request).

Heterogeneous Treatment Effects

Although we are unable to detect an overall treatment effect from introducing ratings to Rialto, it is possible that there *is* a treatment effect for certain subsets of the data. For instance, the impact of ratings on Rialto may evolve over time, or ratings may have a heterogeneous effect on the market dynamics for different item categories.

Heterogeneity Over Time While the specification in Equation 1 measures the overall effect of introducing ratings on market outcomes, it does not capture any heterogeneity in that treatment effect over time. We might expect that the effect of ratings varies as ratings have been live for longer in a given state, since only a few shoppers and sellers have had an opportunity to rate one another immediately following launch. However, as the number of ratings accumulates, the impact may become more salient.

To test for heterogeneous treatment effects over time, we estimate the following difference-in-differences model:

$$Y_{st} = \alpha_s + \gamma_t + \sum_j \delta_j \mathbf{1}\left(\text{ratings}_{st} \in j\right) + \epsilon_{st}, \quad (2)$$

where Y_{st} is the dependent variable in state s at time t, ratings_{st} is the number of days that ratings have been live in a given state, $j = \{<2 \text{ wks}, 2\text{-}4 \text{ wks}, 4\text{-}6 \text{ wks}, 6\text{-}8 \text{ wks}, 8+ \text{ wks}\}$ is a series of time intervals splitting the post-treatment period into 2-week long chunks, and the indicator variable $\mathbf{1}\left(\text{ratings}_{st} \in j\right)$ is 1 when ratings have been live in state s for a length of time in j. α_s is a state-level fixed effect, γ_t is a day-level fixed effect, and ϵ_{st} is the error.

In general, our experiment does not have sufficient statistical power to precisely estimate the heterogeneous treatment effect of ratings over time. However, Figure 2 shows that for all six activity- and transaction-related outcomes,

	Dependent variable:					
		H			$H*$	
	beta	*OLS*	*logistic*	*beta*	*OLS*	*logistic*
	(1)	(2)	(3)	(4)	(5)	(6)
Ratings	−0.011	−0.00001	−0.007	−0.010	−0.00001	−0.007
	(0.044)	(0.00002)	(0.046)	(0.042)	(0.00002)	(0.043)
Observations	7,301	7,301	7,301	7,301	7,301	7,301
State-level F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Time F.E.	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.985	0.751		0.987	0.750	
AIC			398.649			398.942

*p<0.1; **p<0.05; ***p<0.01

Note: Constant term omitted to maintain the confidentiality of financially sensitive Rialto data.

Table 6: Market concentration regression results (7 day window)

the treatment effect point estimates exhibit an increasing pattern the longer ratings have been live in a given state. While these trends are suggestive of potential positive long-term effects, they do not reach conventional levels of statistical significance. In contrast, no clear temporal pattern emerges in the effect of ratings on either the number of reported shoppers or reported sellers, consistent with the limited number of reports in our sample. Analyses of price distribution and market concentration outcomes (results available upon request) exhibit similar patterns of increasing point estimates over time, though these effects likewise remain statistically insignificant.

Heterogeneity Across Categories As described earlier, each item posted to Rialto is assigned one of thirty-two categories, with Figure 1 showing the distribution of these categories by median price and posting volume. We test for heterogeneous treatment effects across the 29 most active categories by estimating the effect of ratings on market activity, price distribution, and market concentration outcomes separately for each category using Equation 1, with standard errors clustered at the state-level using the wild cluster bootstrap.[18] Across all category-metric combinations, few treatment effects reach conventional significance levels, consistent with what would be expected under multiple testing rather than true heterogeneous effects. While we do not report these estimates due to space constraints, detailed category-specific results are available upon request.

Discussion

We find that, on average, centralized reputation did not have a statistically significant impact on Rialto user behavior or market structure in the six months following its introduction. This finding stands in stark contrast to previous studies of the effect of centralized reputation systems (Cai et al. 2014; Bolton, Katok, and Ockenfels 2004), and suggests that the social network context may itself substitute for formal ratings: a seller's profile, mutual connections, group memberships, and transaction history are all signals buyers can potentially use to assess trustworthiness without relying on centralized mechanisms.

The possibility that social network embeddedness can substitute for centralized reputation has major implications for online marketplace designers. Previous work has shown that information collected by reputation systems can be biased due to a number of factors, including herding (Muchnik, Aral, and Taylor 2013), social interaction between buyers and sellers (Fradkin, Grewal, and Holtz 2021), strategic reviewing behavior (Nosko and Tadelis 2015), or long-term "reputation inflation" (Filippas, Horton, and Golden 2022). Furthermore, previous work (Granovetter 1985; Bolton, Katok, and Ockenfels 2004) suggests that information "flows" matter, and that people have greater confidence in information coming from trusted informants or their own dealings. Given the difficulty of designing an effective reputation system, marketplace designers may find it attractive to augment their online marketplaces with social network data, or to embed their online marketplaces in social networks. This may be particularly effective in the early days of an online marketplace, during which there may not be enough transaction volume for a reputation system to be effective.

In choosing between investments in reputation systems and investments in augmenting marketplaces with social network data, there are many factors for platform designers to consider. Augmenting marketplaces with social network information may increase trust and allows marketplaces to avoid many of the biases that are common in online reviews. Furthermore, incorporating social network information into online markets may create additional network effects that help a marketplace grow. However, social network-based trust mechanisms raise important concerns about equity and broader applicability. Users with poor network positions (e.g., low centrality, low degree) may find it difficult to succeed in such marketplaces, and social network-based trust may systematically disadvantage users with limited so-

[18] When estimating category-specific treatment effects for price distribution and market concentration outcomes, we omit state-day-category observations with no posted items.

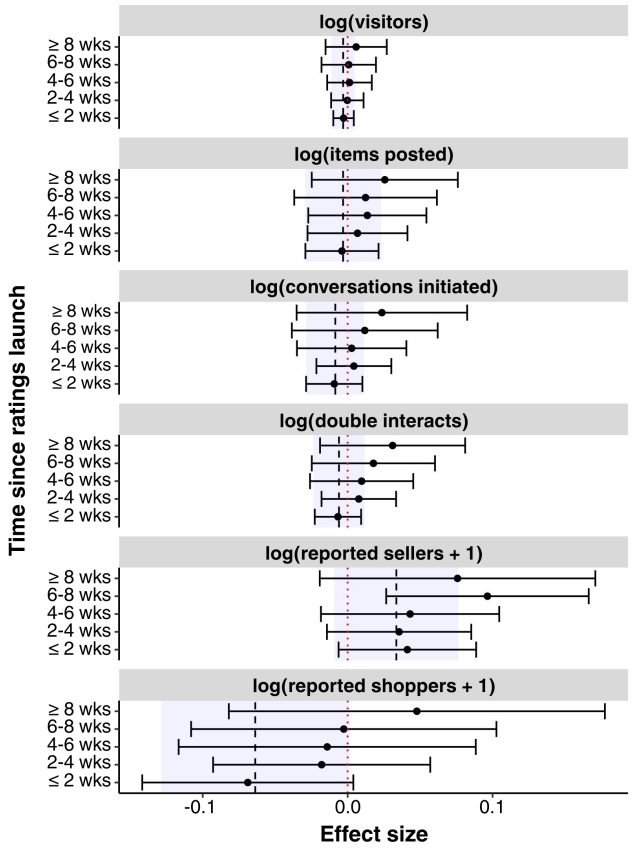

Figure 2: The overall and time-varying effect of ratings on activity outcomes. The black dashed line is the overall treatment effect, and the shaded blue area represents the wild cluster bootstrap 95% confidence interval on the overall treatment effect estimate.

cial connections or from marginalized communities, potentially exacerbating existing digital divides. Additionally, it remains unclear from existing research whether a shopper or seller's network attributes are actually predictive of their quality as a transaction partner. While this question warrants further research, a combination of online reputation and social network embeddedness may allow platforms to get the best of both worlds.

Given these complexities, we caution that our findings about the limited impact of reputation systems should not be misused to justify avoiding such systems in contexts where they remain essential. Our results are specific to socially embedded marketplaces and should not be misinterpreted as justification for avoiding formal safeguards in anonymous marketplaces or those serving diverse user bases without strong social connectivity. Indeed, centralized reputation is likely to remain particularly valuable in settings where users lack pre-existing social connections, where transactions involve high stakes or repeated interactions with strangers, and/or where users can easily create pseudonymous accounts.

Of course, there are alternate explanations for why Ri-alto's reputation system did not, on average, have an impact. One possibility is that not enough shoppers and sellers left ratings during our experiment for Rialto to have a large effect. Figure 4 in the appendix shows the percentage of the time that shoppers and sellers left ratings. Neither rate rises above 3% for the duration of our experiment.[19] It is well-documented that reviews are a public good that often go underprovisioned (Avery, Resnick, and Zeckhauser 1999; Resnick and Zeckhauser 2002; Bolton, Katok, and Ockenfels 2004). In the case of Rialto, it is possible that specific design decisions (e.g., the entry points through which users can rate one another, the ability for seller's to hide their ratings) exacerbated this tendency. After a series of changes to the product's design, ratings may have a more pronounced effect on Rialto.

It is also possible that the introduction of centralized reputation had a long-run effect on Rialto, but that there is no short-run effect. While they are not statistically significant, the point estimates in our analysis of time-varying treatment effect heterogeneity suggest this may be the case. For almost all metrics, the effect of ratings seems to be increasingly positive the longer ratings have been live on Rialto. Given the relatively short duration of our experiment (and the aforementioned underprovision of ratings), our experiment may not have run long enough to detect these long-run effects with statistical significance.

Although our findings may not generalize to all marketplace contexts, they contribute empirical evidence for an underexplored possibility: that informal social networks can substitute for formal reputation systems. Although our main result is a null result, we argue that it represents a significant contribution to the existing research on both reputation systems and socially embedded commerce. Throughout the history of both the physical and social sciences, null results have played an important role in inspiring new theories (Michelson and Morley 1887), refuting commonly held beliefs (Banerjee et al. 2015), and informing policy debates (Abdulkadiroğlu, Angrist, and Pathak 2014). More research should be done to better understand how social network structure and centralized reputation can substitute for and complement one another, the extent to which the impact of reputation systems is dependent on the characteristics of the goods being sold, how the impact of a reputation system evolves over time, and how different design choices can affect a reputation system's efficacy.

Acknowledgments

The authors would like to thank Alhad Purnapatre, Desi Wang, Francesco Fogu, Gulin Yilmaz, Jason Feng, Yatharth Saraf, Erica Klein, Vivek Sharma, Mingchen Zhao, Lauri Kanerva, Shawna Jemison, Reza Khosravani, Rajani Aswani, Kevin Mei, Helen Wei Zeng, and Lada Adamic for their help with this project. We are grateful to Dean Eckles and John Horton for their comments and feedback, and to Jonathan Hazell for useful conversations regarding method-

[19] As previously mentioned, on Rialto shoppers and sellers can rate most counterparties they interact with. This makes it difficult to compare this 3% rate to review rates on other platforms.

ology. We also thank Brian Karrer for early conversations regarding this project. We are especially grateful to Mike Bailey and Steve Tadelis for their encouragement and support in bringing this work to publication. This work was deemed exempt status by the MIT Committee On the Use of Humans as Experimental Subjects under Protocol #1803264564. Part of this work was conducted while Holtz was an intern and, later, a contractor at Facebook, Inc., and while Dow was a full-time employee at Facebook, Inc.

References

Abdulkadiroğlu, A.; Angrist, J.; and Pathak, P. 2014. The elite illusion: Achievement effects at Boston and New York exam schools. *Econometrica*, 82(1): 137–196.

Akerlof, G. A. 1978. The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in Economics*, 235–251. Elsevier.

Avery, C.; Resnick, P.; and Zeckhauser, R. 1999. The market for evaluations. *American Economic Review*, 89(3): 564–584.

Banerjee, A.; Duflo, E.; Glennerster, R.; and Kinnan, C. 2015. The miracle of microfinance? Evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, 7(1): 22–53.

Bapna, R.; Ramaprasad, J.; Shmueli, G.; and Umyarov, A. 2016. One-way mirrors in online dating: A randomized field experiment. *Management Science*, 62(11): 3100–3122.

Bar-Isaac, H.; Tadelis, S.; et al. 2008. Seller reputation. *Foundations and Trends® in Microeconomics*, 4(4): 273–351.

Barrios, T.; Diamond, R.; Imbens, G. W.; and Kolesár, M. 2012. Clustering, spatial correlations, and randomization inference. *Journal of the American Statistical Association*, 107(498): 578–591.

Bertrand, M.; Duflo, E.; and Mullainathan, S. 2004. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1): 249–275.

Bohnet, I.; and Huck, S. 2004. Repetition and reputation: Implications for trust and trustworthiness when institutions change. *American economic review*, 94(2): 362–366.

Bolton, G.; Greiner, B.; and Ockenfels, A. 2013. Engineering trust: reciprocity in the production of reputation information. *Management science*, 59(2): 265–285.

Bolton, G. E.; Katok, E.; and Ockenfels, A. 2004. How effective are electronic reputation mechanisms? An experimental investigation. *Management science*, 50(11): 1587–1602.

Burbidge, J. B.; Magee, L.; and Robb, A. L. 1988. Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association*, 83(401): 123–127.

Cabral, L.; and Hortacsu, A. 2010. The dynamics of seller reputation: Evidence from eBay. *The Journal of Industrial Economics*, 58(1): 54–78.

Cai, H.; Jin, G. Z.; Liu, C.; and Zhou, L.-a. 2014. Seller reputation: From word-of-mouth to centralized feedback. *International Journal of Industrial Organization*, 34: 51–65.

Callaway, B.; and Sant'Anna, P. H. C. 2021. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2): 200–230.

Cameron, A. C.; Gelbach, J. B.; and Miller, D. L. 2008. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3): 414–427.

Coleman, J. S. 1988. Social capital in the creation of human capital. *American Journal of Sociology*, 94: S95–S120.

de Chaisemartin, C.; and D'Haultfœuille, X. 2020. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9): 2964–2996.

Dellarocas, C. 2003. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science*, 49(10): 1407–1424.

Ferrari, S.; and Cribari-Neto, F. 2004. Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7): 799–815.

Filippas, A.; Horton, J. J.; and Golden, J. M. 2022. Reputation inflation. *Marketing Science*, 41(4): 733–752.

Fisher, R. A. 2006. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Fradkin, A.; Grewal, E.; and Holtz, D. 2021. Reciprocity and unveiling in two-sided reputation systems: Evidence from an experiment on Airbnb. *Marketing Science*, 40(6): 1013–1029.

Friedman, E. J.; and Resnick, P. 2001. The social cost of cheap pseudonyms. *Journal of Economics & Management Strategy*, 10(2): 173–199.

Garg, N.; and Johari, R. 2021. Designing informative rating systems: Evidence from an online labor market. *Manufacturing & Service Operations Management*, 23(3): 589–605.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Goodman-Bacon, A. 2021. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2): 254–277.

Granovetter, M. 1985. Economic action and social structure: The problem of embeddedness. *American journal of sociology*, 91(3): 481–510.

Holtz, D.; MacLean, D. L.; and Aral, S. 2017. Social Structure and Trust in Massive Digital Markets. In *ICIS 2017 Proceedings*, 13.

Jin, G. Z.; and Kato, A. 2006. Price, quality, and reputation: Evidence from an online field experiment. *The RAND Journal of Economics*, 37(4): 983–1005.

Klein, T. J.; Lambertz, C.; and Stahl, K. O. 2016. Market transparency, adverse selection, and moral hazard. *Journal of political economy*, 124(6): 1677–1713.

Kreps, D. M. 1996. Corporate culture and economic theory. *Firms, Organizations and Contracts, Oxford University Press, Oxford*, 221–275.

Livingston, J. A. 2005. How valuable is a good reputation? A sample selection model of internet auctions. *Review of Economics and Statistics*, 87(3): 453–465.

MacKinnon, J. G.; and Magee, L. 1990. Transforming the dependent variable in regression models. *International Economic Review*, 315–339.

Michelson, A. A.; and Morley, E. W. 1887. On the Relative Motion of the Earth and of the Luminiferous Ether. *Sidereal Messenger, vol. 6, pp. 306-310*, 6: 306–310.

Moore, R. T. 2012. Multivariate continuous blocking to improve political science experiments. *Political Analysis*, 20(4): 460–479.

Muchnik, L.; Aral, S.; and Taylor, S. J. 2013. Social influence bias: A randomized experiment. *Science*, 341(6146): 647–651.

Nosko, C.; and Tadelis, S. 2015. The limits of reputation in platform markets: An empirical analysis and field experiment. Technical report, National Bureau of Economic Research.

Resnick, P.; and Zeckhauser, R. 2002. Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. In *The Economics of the Internet and E-commerce*, 127–157. Emerald Group Publishing Limited.

Resnick, P.; Zeckhauser, R.; Swanson, J.; and Lockwood, K. 2006. The value of reputation on eBay: A controlled experiment. *Experimental economics*, 9(2): 79–101.

Rosenbaum, P. R. 1984. Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, 79(387): 565–574.

Rosenbaum, P. R. 2002. *Observational Studies*. New York: Springer, 2nd edition.

Silva, J. S.; and Tenreyro, S. 2006. The log of gravity. *The Review of Economics and statistics*, 88(4): 641–658.

Tadelis, S. 2016. Reputation and feedback systems in online platform markets. *Annual Review of Economics*, 8: 321–340.

Weick, K. E. 1979. Cognitive processes in organizations. *Research in organizational behavior*, 1(1): 41–74.

Wooldridge, J. M. 2006. *Introduction to Econometrics: A Modern Approach*. South-Western.

Zervas, G.; Proserpio, D.; and Byers, J. W. 2021. A first look at online reputation on Airbnb, where every stay is above average. *Marketing Letters*, 32(1): 1–16.

Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, the research examines marketplace design in ways that could improve online commerce while using anonymized platform data with appropriate privacy protections.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the abstract and introduction clearly state that we find null effects of reputation systems in socially embedded marketplaces, which matches our empirical findings.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we use a randomized staggered rollout with matched-pairs design and difference-in-differences analysis, which is appropriate for establishing causal effects of reputation system introduction.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we discuss several potential data artifacts: data is scaled by an unreported constant for confidentiality (preserving relationships), most sellers are casual creating skewed posting distributions, offline transactions make "sold" status an imperfect measure, and our geographic sample excludes Washington state where ratings were already available.
   (e) Did you describe the limitations of your work? Yes, we discuss low rating adoption rates ($< 3\%$), potential design issues, and short-term vs. long-term effects as limitations of our study.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, we discuss in the Discussion section how social network-based trust might disadvantage users with poor network positions and acknowledge risks of overgeneralization to inappropriate contexts.
   (g) Did you discuss any potential misuse of your work? Yes, we explicitly caution that our findings about the limited impact of reputation systems should not be misused to justify avoiding such systems in non-social marketplace contexts where they may remain essential for establishing trust between users.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we implemented several protective measures: anonymized all platform and company names, conducted analysis on de-identified anonymous data, scaled data by unreported constants to protect commercially sensitive information, conducted all analysis on secure company company systems with appropriate access controls, and obtained IRB exemption for the use of anonymized behavioral data.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes.
   (b) Have you provided justifications for all theoretical results? Yes, we provide and/or describe extensive robustness checks including inverse hyperbolic sine

transformations, quasi-Poisson models, randomization inference analysis, and alternative smoothing windows. We also show a parallel trends figure that demonstrates that our identifying assumptions hold.

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes.

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, we extensively discuss alternative explanations including: low rating adoption rates, suboptimal system design, and differences between short-term and long-term effects. We also test for heterogeneous effects across time periods and product categories.

(e) Did you address potential biases or limitations in your theoretical framework? NA - This study does not introduce a new theoretical framework but rather empirically tests the validity of existing theoretical frameworks regarding social embeddedness and reputation systems in online marketplaces.

(f) Have you related your theoretical results to the existing literature in social science? Yes, we extensively relate our findings to prior work on reputation systems (Cai et al. 2014; Bolton, Katok, and Ockenfels 2004), social embeddedness theory (Granovetter 1985), and empirical studies of social commerce (Holtz, MacLean, and Aral 2017). We discuss how our null findings contrast with previous reputation system studies and align with social embeddedness literature.

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, we discuss implications for marketplace designers considering whether to invest in reputation systems versus social network features, acknowledge potential negative impacts on users with poor network positions, and outline several areas for future research including boundary conditions and design optimization.

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? NA

(b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? NA

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA - This study analyzes behavioral data from users naturally engaging with a platform feature rollout, not direct participation with researcher-provided instructions.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, we mention in a footnote that this research was deemed exempt by an Institutional Review Board as it involves analysis of anonymized behavioral data from a staggered rollout conducted for business purposes, indicating that potential participant risks were assessed and found to be minimal.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA - No participants were recruited or compensated; the study observes natural user behavior during a platform feature rollout.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, we note in a footnote that data analysis was conducted on company systems with deidentified user data and appropriate access controls.

Analysis Using Randomization Inference

One potential issue with the main text analysis is that it does not properly account for the matched pairs design of our experiment. This could lead to overly conservative standard errors and a failure to report statistically significant treatment effects. To verify this is not the case, in this section

we re-analyze our experimental data using Fisherian conditional randomization inference.[20] We condition our randomization inference on the matched state pairs obtained via the matching procedure described in Section . We test the sharp null hypothesis of no treatment effect (i.e., $Y_{st}(\text{ratings}_{st} > 0) = Y_{st}(\text{ratings}_{st} = 0) \,\forall\, s, t$) and calculate p-values using a Monte Carlo approach ($n = 1,000$).

Table 7: State pairs, Mahalanobis distances, and treatment assignments. Ratings were introduced to treatment states through a staggered rollout lasting from December 5, 2017 to February 21, 2018. Ratings were introduced to control states on February 22, 2018.

Treatment	Control	Mahal. Dist.
RI	HI	0.53
MA	CT	0.63
ND	SD	0.64
NM	AK	0.83
NH	VT	0.90
KS	NE	1.08
DE	NV	1.16
ID	WY	1.44
MS	SC	1.46
MT	OR	1.77
TN	GA	1.93
AR	LA	2.12
MD	NC	2.14
IL	IN	2.15
PA	MI	2.33
CO	UT	2.46
ME	AZ	2.60
OK	AL	2.83
MN	WI	3.46
KY	MO	3.75
WV	OH	4.12
NJ	CA	4.77
FL	VA	5.03
NY	IA	6.62
TX	N/A	N/A

We also use Fisherian conditional randomization inference to obtain a 95% confidence interval for the treatment effect. We assume the following model for each state-day's potential outcomes both with and without ratings:

$$Y_{st}(\text{ratings}_{st} > 0) = \tau Y_{st}(\text{ratings}_{st} = 0), \tag{3}$$

i.e., the treatment has a constant, multiplicative effect for all state-days.[21] We calculate a p-value using Fisherian inference for each value of τ on a grid around 1.00, with a spacing of 0.001 between grid points. The set of τ for which $p > 0.05$ provide a 95% confidence interval. Table 8 shows the p-values calculated for each outcome using Fisherian conditional randomization inference, as well as the lower and upper bounds of the treatment effect confidence interval under the constant, multiplicative effect model.[22] The effect of ratings on each of our outcomes remains statistically insignificant, with the null results for non-reporting-related metrics precisely estimated.

[20]For a detailed explanation of Fisherian randomization inference, we refer the reader to (Fisher 2006; Rosenbaum 1984, 2002) and (Barrios et al. 2012).

[21]Under this constant, multiplicative treatment effect model of potential outcomes, our sharp null hypothesis corresponds to the case where $\tau = 1$.

Robustness Checks

Re-estimation With Inverse Hyperbolic Sine Transformation

In addition to log-transformed outcomes, we re-estimate the effect of introducing ratings on the inverse hyperbolic sine transformation (Burbidge, Magee, and Robb 1988; MacKinnon and Magee 1990) of number of Rialto visitors, number of items posted to Rialto, number of product detail page views, number of conversations initiated, number of double interacts, number of items marked as sold, number of reported sellers, and number of reported shoppers. The inverse hyperbolic sine transformation is expressed as

$$\text{asinh}\,(y) = \log\left(y + \left(y^2 + 1\right)^{\frac{1}{2}}\right). \tag{4}$$

Interpretation of regression coefficients for inverse hyperbolic sine-transformed outcomes is almost identical to interpretation of regression coefficients for log-transformed outcomes, with the added benefit that inverse hyperbolic sine is defined at zero (meaning it is not necessary to add a uniform constant to variables for which zeros occur).

Qualitatively, the results we obtain when estimating our linear model on inverse hyperbolic sine transformed outcomes are extremely similar to those obtained from a linear model estimated on log-transformed outcomes. We are not able to detect a statistically significant effect of ratings on any of our outcome metrics. This finding tends to hold for almost all specific category-outcome pairs as well.

Re-estimation With Quasi-Poisson Model

We also re-estimate the effect of introducing ratings on number of Rialto visitors, number of items posted to Rialto, number of product detail page views, number of conversations initiated, number of double interacts, number of items marked as sold, number of reported sellers, and number of reported shoppers using a quasi-Poisson regression model (Wooldridge 2006; Silva and Tenreyro 2006). The quasi-Poisson regression approach models the probability that $y = h$ conditional on $\mathbf{x}$ as

$$P(y = h|\mathbf{x}) = \frac{[\exp(\mathbf{x}\beta)]^h \exp(-\exp(\mathbf{x}\beta))}{h!}, \tag{5}$$

and assumes that the variance is proportional to the mean,

$$\text{Var}\,(y|\mathbf{x}) = \sigma^2 \mathbb{E}\,(y|\mathbf{x}) \tag{6}$$

[22]We do not conduct Fisherian conditional randomization inference to re-estimate the effect of ratings on market concentration due to computational cost.

Table 8: p-values and confidence interval lower and upper bounds obtained through conditional Fisherian randomization inference

Outcome	p-value	τ lower bound	τ upper bound
log(visitors)	0.12	99.0%	100.0%
log(posts)	0.46	97.7%	101.0%
log(product detail page views)	0.34	98.3%	100.6%
log(conversations initiated)	0.23	97.9%	100.5%
log(double interacts)	0.37	98.1%	100.6%
log(items marked as sold)	0.06	97.0%	100.1%
log(reported shoppers + 1)	0.69	96.6%	105.0%
log(reported sellers + 1)	0.10	95.0%	100.6%
Median price	0.04	99.9%	104.3%
90%/10%	0.24	98.0%	107.5%
90%/50%	0.63	96.7%	105.3%
50%/10%	0.31	99.5%	102.1%

where σ^2 is the dispersion parameter that allows for overdispersion ($\sigma^2 > 1$) or underdispersion ($\sigma^2 < 1$), unlike standard Poisson regression which assumes $\sigma^2 = 1$.

Qualitatively, the results we obtain from quasi-Poisson regression are extremely similar to those obtained from a linear model estimated on log-transformed outcomes. We are not able to detect a statistically significant effect of ratings on any of our outcome metrics. This finding tends to hold for almost all specific category-outcome pairs as well.

Market Concentration Measures: Mathematical Definitions

To quantify market concentration, we first define s_{sjt}, the share of items posted in state s on day t coming from seller j, as:

$$s_{sjt} = \frac{\sum_{\text{items posts}_{st}} \mathbf{1}\left(\text{seller} = j\right)}{\sum_{\text{items posts}_{st}} 1}. \tag{7}$$

We can then define two metrics: the Herfindahl index:

$$H_{st} = \sum_{j \in st} s_{sjt}^2, \tag{8}$$

and the normalized Herfindahl index:

$$H^*_{st} = \frac{H_{st} - \frac{1}{N_{st}}}{1 - \frac{1}{N_{st}}} \quad \text{for } N_{st} > 1 \tag{9}$$
$$H^*_{st} = 1 \quad \text{for } N_{st} = 1$$

To account for the sporadic nature of sellers' posting activity, we compute smoothed versions using a smoothed version of s_{sjt}:

$$s^{smooth}_{sjt} = \frac{\sum_{T=t-7}^{T=t+7} \sum_{\text{items posts}_{sT}} \frac{1}{2}^{|T-t|} \mathbf{1}\left(\text{seller} = j\right)}{\sum_{T=t-7}^{T=t+7} \sum_{\text{items posts}_{sT}} \frac{1}{2}^{|T-t|}}, \tag{10}$$

which takes into account exponentially decayed past and future market participation with a window of seven days. We then calculate both the Herfindahl index and normalized Herfindahl index using the smoothed version of s_{sjt}.

Additional Figures and Tables

Table 7 presents the complete list of state pairs used in our matched-pairs experimental design, showing the Mahalanobis distances between paired states and their treatment/control assignments. The distances range from 0.53 (Rhode Island and Hawaii) to 6.62 (New York and Iowa), with Texas remaining unmatched due to the odd number of states. Table 8 reports the results of our Fisherian conditional randomization inference analysis, which accounts for the matched-pairs design and tests the sharp null hypothesis of no treatment effect. The p-values and confidence intervals confirm that our null results remain statistically insignificant when properly accounting for the experimental design.

Figure 3 demonstrates the parallel trends assumption underlying our difference-in-differences analysis, showing that treatment and control states exhibited similar trajectories in items posted per state-day before the December 5, 2017 rollout began. Figure 4 provides detailed descriptive statistics on rating behavior throughout the experiment period, revealing that both shopper and seller rating rates remained extremely low (below 3%) despite multiple entry points for users to provide feedback, while positive ratings consistently comprised over 80% of all ratings given.

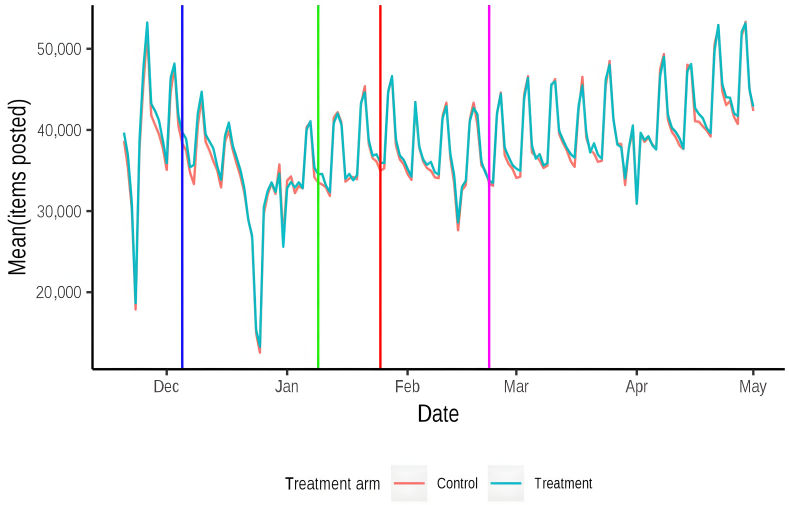

Figure 3: The average number of items posted per state-day in the control and treatment. The blue, green, red, and magenta vertical lines correspond to the dates on which ratings were introduced to new states. y-axis values have been scaled by an unreported constant to maintain the confidentiality of financially sensitive Rialto data.

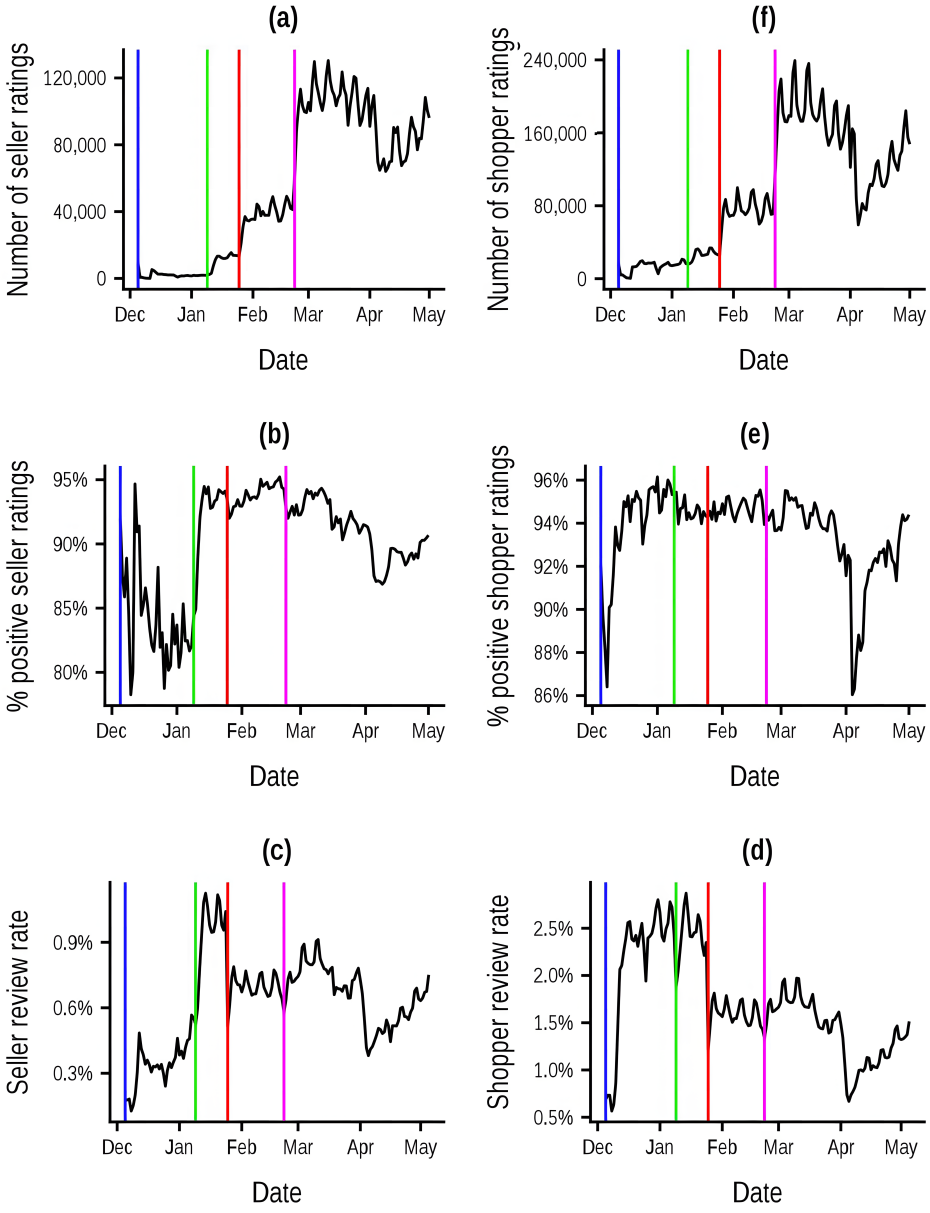

Figure 4: Counterclockwise from the top left, this figure shows (a) the number of seller ratings per day over time (b) the percentage of seller ratings per day that were positive (c) the rate at which shoppers rated sellers (ratings / conversations with a reply) (d) the rate at which sellers rated shoppers (ratings / conversations) (e) the percentage of shopper ratings per day that were positive (f) the number of shopper ratings per day over time. y-axis values in (a) and (b) have been scaled by an unreported constant to maintain the confidentiality of financially sensitive Rialto data.