Resisting Reputation Manipulation in Marketplaces - Business Analysis

Business Analysis

Introduction

Online activity is currently reshaping the shopping experience. As the rise of online shopping creates many new business opportunities and increases consumer choices, a host of new questions and problems also arises. One big question, for example, is how can the average consumer wade through the overwhelming mass of data to find products that best fit his or her needs at the best price? To address this question many shopping sites use product ratings via user reviews, and recommendation engines. Yet this leads us to another pertinent question: how trustworthy are these ratings and recommendation systems? As described in our SingularityNET market research (posted by Tim Richmond, June 21st at https://blog.singularitynet.io/understanding-the-user-behavior-of-digital-marketplaces-355cc0f59af5), these systems often create incentives for sellers to create fake reviews and ratings in order to boost sales, including financial incentives to otherwise honest buyers to uprate or downrate products. In this article, we build upon the results of our previous market research to provide additional detailed analyses and improvements to our current reputation system to better handle these scam campaigns.

What is the problem?

When online marketplaces were introduced, fake reviews and rating manipulations were not seen as a problem. Sellers quickly noticed that positive ratings and reviews massively improved product sales due to higher rankings on recommendation engines. Reputable suppliers responded by asking customers to leave reviews and feedback for future product improvements. Some disreputable suppliers, on the other hand, began gaming the system in order to artificially inflate ratings by, for example, paying individuals to provide fake product reviews and ratings. While it is hard to estimate when first fake reviews occurred, this started to happen on a relatively large scale around a decade ago[1].

Online marketplaces try to catch fake reviewers by using machine learning algorithms for anomaly detection. They also develop models to minimize the effects of manipulation. Despite those measures fake reviews remain a huge problem on online marketplaces with an estimated

[1] A quick view on Google trends indicates that it has been on the rise since 2009 https://trends.google.com/trends/explore?date=all&geo=US&q=fake%20reviews,amazon%20fake%20review

15-20% of reviews being fake. Around 10% of the reviews of the biggest online shop, Amazon, are fake[2] (different sources estimate different numbers, but this seems to be the average.) While suppliers, customers and researchers are all bringing attention to this problem, large marketplaces find the problem difficult to admit to – Amazon has publicly said that only about one percent of the reviews on its platform are fake[3] – a figure with which other sources disagree.

Manipulated reviews can affect overall ratings and therefore rankings of specific products. Marketplaces often rely on a high frequency of reviews from real buyers to offset possible rating manipulation. In most marketplaces, around 15-30% of purchasers leave reviews or ratings, thus requiring scam sellers to leave many fake reviews in order to maintain a good rating.

Actual percentages of fake reviews depend on several factors including the size of the marketplace, the nature of product experiences, and the ease or difficulty of the process for leaving reviews. It is estimated that only around 2-5% of users leave reviews on Amazon.com since users must be registered and write out entire reviews. Such free-answer reviews require considerably more investment from customers than do simple star-ratings without a review-writing requirement. Furthermore, people satisfied with products tend not to exert effort writing reviews. On the platforms where leaving reviews is relatively difficult (such as Amazon), ratings are often distributed differently than on those platforms upon which leaving reviews is easier. An article in the New York Times suggests that 4.3 stars on Amazon "might be the result of a hard-fought and expensive campaign to climb to the first page of the search results" while the same 4.3 stars on Uber "might mean a driver is at risk of getting booted from the system". The reasoning behind this assertion is that the ease with which customers are able to provide ratings with a single tap on Uber, compared to the daunting requirement of typing an entire review on Amazon, should result in higher average ratings for Uber.

What We Did

In our previous work, we already developed a reputation algorithm that is designed to tackle fake reviews and ratings on online marketplaces. In the continuation of our work, we made a detailed market research on fake reviews around different marketplaces and we specifically focused on the Amazon case. Our market research and detailed analyses derived from it was then used to create a large scale simulation. In this simulation we reproduced both the quantity of manipulated reviews and financial rewards that scamming suppliers received to better understand their motivations and the effects of different market conditions.

Since fake raters work to conceal their behavior, it is easier to estimate how various factors, such as ratings distributions and other ratings and review patterns, affect the total amount of scam than to identify individual fake reviewers. Due to this difficulty in detecting specific bad

---

[2] https://www.buzzfeednews.com/article/nicolenguyen/amazon-fake-review-problem

[3] https://finance.yahoo.com/news/rise-fake-amazon-reviews-spot-175430368.html?guccounter=1&guce_referrer=aHR0cHM6Ly9kdWNrZHVja2dvLmNvbS8&guce_referrer_sig=AQAAAH3Frj6oo2Caa1C_WU70WDLauDjGZHKJqQ5Legly3OyM1PymT9UrgfM4eevolWKjkgr_t8Hn3Y48gj2VUzlOl1qgndUL9EDFOUHsTdYnwd_3P4IuFsi6hi0FliGQAODtftcMrqb1PePiCFGsrqPkSO95je4ETbo5MsV53FvoLZUS

actors, our system is instead designed to minimize the total negative effects of scamming behavior by minimizing incentives and maximizing costs to potential scammers.

We set the parameter values for our simulation based on our prior market research as follows;

- We set the amount of fake reviewers to be roughly 20%. The amount of fake reviews is towards the higher end - most marketplaces have the average of fake reviews between 15-20%.
- The chance of a customer leaving a review is set to be at 30% for all of the reviewers. Since scammers will leave reviews and therefore ratings at 100% rate (so, the scam is worth doing), real reviewers leave ratings a bit less often in practice.
- Probability of leaving a review is different for different marketplaces. While it stands at around 2-5% for Amazon, it is much higher on other marketplaces. Indeed, in the marketplace in which our system works best, we do not necessarily have to leave a review with a rating. In such a marketplace, it takes substantially less effort for a user to leave a review than in the example of Amazon. This is why we decided for a bit higher probability of leaving a review - in our case it is 30% (we also did experiments with 70%).
- The distribution of prices of different products, distribution of ratings and price of scam were determined in market research and we took those findings into account in our simulation. We determined that each fake (“five stars”) review costs the seller a refund of the fake sale plus the production cost of the product..
- We determined the prices of the goods on the market (including the ones that are manipulated) to be in range \$5-\$500.,
- Each seller may have an average of 100 products - all within the same price range. So buyers are rating and buying specific products but the cash flows to the suppliers of the products who sponsors all fake reviews across their product lines.
- Organic buys (that is the buys made by honest buyers) are made in natural timing patterns.
- In our simulation the sellers and buyers are not overlapping.
- Fake sellers have their accounts periodically banned for reputation gaming and are forced to create new accounts with fake identities in order to keep sales.
- All fake reviews/ratings were set to be at the 5 star level. It makes sense for them to do the positive reputation gaming, so they maximize their ratings.
- In simulation the fake “sponsored” consumers may perform “organic” buys in natural consumption patterns and providing natural (not “all 5 stars” necessarily) ratings to other products occasionally.
- In our simulation, buyers have memory, that is if a buyer *a* has bought something from seller *b*, and the quality has turned to be low, they will not repeat a purchase form the same supplier since they were disappointed with the product. The opposite holds true for high quality products - buyers are more likely to repeat a purchase of good products.
- We enabled scamming sellers to replace a product that is not doing well, or even change its entire line of products and identity if simple product replacement does not improve its situation.

- When comparing to Amazon, we found that there are also roughly the same number of products as there are buyers, so the ratios are close to what one can expect on a popular marketplace[4,5].

We improved our liquid rank reputation algorithm by adding additional parameters which we then adjusted to optimally suppress scamming effects. We also developed several metrics in order to analyze algorithm performance. The metrics we use are:

- **OMU: Organic market utility** - This is the amount of buys that legitimate buyers made from legitimate sellers based on ratings, divided by the maximum amount of buys possible in the absence of ratings. We wish to maximize this metric.
- **LTS: Loss To Scam** - This is defined as the fraction of all spendings made by honest buyers paid to dishonest sellers. This metric shows what proportion of money spent by honest buyers is spent on the products offered by dishonest sellers, so we want to minimise it.
- **BSL: Buyer Satisfaction Loss** - This is calculated as the average of expected dissatisfaction on behalf of honest buyers weighted by financial values. It is the sum of overall loss (values of individual organic buys of products times expected dissatisfaction with each product) divided by overall spending (the total value of organic buys). Here, the expected dissatisfaction is calculated as 1 minus the known product quality, where quality is measured as a value between 0 (low quality) and 1 (high quality.) We want to minimize BSL, that so that honest buyers get the highest quality products possible, with account to the product cost.
- **SGP: Seller Gaming Profit** - This is the sum of scam income (organic buys of products times dissatisfaction with each product) divided by the cost of scam (sum of all sponsored buys by dishonest buyers together with their commision rate). Low SGP values correspond to low profits by scamming sellers from honest buyers, so we wish to minimize it.

We also added new extensions to the existing liquid rank reputation algorithm, called "Anti-biased", "Predictive" and "Vendor Impact", as briefly explained below.

- "**Anti-biased**" feature takes into account buyers' ratings histories attempting to solve the rater bias problem. We know that in real marketplaces buyers are often biased. Some buyers tend to always post 5-star review or in general they are more satisfied with products than average. The other buyers however, are feeling unhappy with everything. Finally, the "gaming-ready" buyers are just making money providing "sponsored ratings" being paid and reimbursed by scamming suppliers for the fake 5-star reviews. The "anti-biased" configuration, when turned on, corrects for biases and adjusts the weights of ratings that those buyers give. It determines the corrections based on past ratings that those buyers have done.

[4] Grey P. (2015): " How Many Products Does Amazon Sell?", https://export-x.com/2015/12/11/how-many-products-does-amazon-sell-2015/

[5] Feedbackexpress (2018): "Amazon Has 1,029,528 New Sellers This Year (Plus Other Stats)", https://www.feedbackexpress.com/amazon-1029528-new-sellers-year-plus-stats/

- The “**Predictive**” configuration of the reputation system also focuses on buyers and checks their review and rating history. This checks what ratings the certain buyer gave to the product and what ranking this product had in the end. If ratings by some buyers tended to predict actual reputation ranks in the past well, then these buyers got higher weight in this are and thus their further ratings are more valuable in the future than a rating of someone whose ratings did not predict later reputation score that well.
- The “**Vendor Impact**” feature applies to the case when buyers are rating not sellers themselves but their products. In such a case the products have the reputation ranks computed for them, and the reputation earned by products is translated to reputation ranks of their vendors. Further on, the reputation of the vendors affects the reputation of their products even if these products are not rated explicitly by any buyer.

We created both non-adaptive and adaptive simulations. In the non-adaptive simulation, scamming vendors sold low-quality products and allocated funds for purchasing ratings from the “gaming-ready” fraction (20%) of buyers until they have created a market for their inferior products. For our non-adaptive simulation we ran simulations based on three different buyer population sizes (100, 500, 1000), three different product selection sizes (100, 500, 1000) and three different lengths of simulation durations (100 days, 500 days, 1000 days.)

In our adaptive simulations, buyers purchased goods and services online in accordance with supplied utility parameters, with purchase priorities increasing with the length of time between purchases. Sellers are assigned to supply these goods in proportion to demand in order to meet needs. The quality of goods and services supplied correlates with an organic quality of the seller. Honest seller ratings are then based upon those organic quality scores along with measures of bias and perception. Consumers remember scamming suppliers from which they have previously purchased and avoid them. They also remember good suppliers, tending to return to them for their needs.

Normal distributions determine buyer and seller traits and Zipfian distributions determine price and utility. Bayesian networks represent the probabilities of various marketplace participant decisions. We use reputation scores and profit amounts to determine how many scam campaigns dishonest buyers will purchase for the products they sell, whether dishonest buyers will change a product that is not doing well, and whether dishonest buyers change their identity. We have one set of run results for 1000 sellers and 1000 buyers and another similar set of results for runs with 10,000 sellers and 1000 buyers. For simplicity, we have set the number of product categories that consumers need to five, one of which the dishonest suppliers engage in selling, and have set each supplier to sell an average of one hundred different products, all of the same price and product category.

Furthermore, for each of those settings we tried the following settings:

- **No reputation system:** Participants make decisions relying only on their own memories and not referring to any reputation system.
- **Regular reputation system**: Does not take into account any factors other than values of ratings that consumers make to suppliers.

- **Weighted reputation system**: Rating values are weighted by costs of transactions that are rated.
- **TOM-based reputation system**: In addition to weighting ratings with financial values per-transaction, weights are based on the rater's time on the market (TOM) as a "proof-of-time". That is, the raters (buyers) are implicitly rated based on how long they have been on the market. So, rating by buyer with a longer history influences reputation of a seller more than the one made by rater with shorter history.
- **SOM-based reputation system**: In addition to weighting ratings with financial values per-transaction, weights are based on rater's spendings on the market (SOM) as a "proof-of-burn" value. That is, the raters (buyers) are implicitly rated based on how much they spend on this market. So, rating by buyer with a lot of spendings influences reputation more than the one made by rater with smaller spendings.
- **Anti-biased reputation system**: A liquid rank reputation system correcting for buyers bias, that is for their average rating when assigning ratings.
- **Predictive reputation system:** Reputation system that uses buyer's predictiveness (that is how good a buyer is at predicting a suppliers future score) in order to determine ratings. We might give different weights to predictiveness, from 0% (we don't take it into account) to 100% where it is very important when calculating ranks.
- **Vendor Impact reputation system**: Reputation system that uses "liquid rank" algorithm to translate reputation of specific products to their vendors and vice versa.

We have used both the adaptive simulation and non-adaptive simulations to explore how different setups of reputation system perform for different markets with the following results.

Our Results

In order for our regular reputation system to work we need as many real buyers with honest feedback as possible. The minimum condition would be for there to be at least as many honest buyers with feedback (left as ratings or reviews) as there are cheaters. If, for example, 30% of the buyers leave feedback and 20% of them are providing fake reviews or ratings, then real consumers represent 80% of the transactions and 24% of the ratings. 24% of the ratings made by honest consumers is more than 20% of cheaters who leave ratings and therefore in such a case, the system works. In our case, we therefore need to have a probability of leaving a review at least 25% for honest buyers. Our results show that using improved settings, we can make a reputation system work with even fewer honest buyers leaving reviews or ratings.

The metric driving the above is called probability of leaving a rating (PLR) and the following table and charts present its impact, based on adaptive simulation for the "Regular" reputation system. We have explored how the financial metrics perform under different market conditions. The results for 1000 buyers and products is shown below, with the chart displaying the performance of the metrics in relationship to PLR. The more users leave ratings, the better the values of the metrics. That intuitively makes sense - since more buyers (legit buyers) leave reviews and the same amount of scam buyers (that is all of them) leave reviews, we get more informed decisions. Note that given our simulation setup, including 20% “gaming-ready” population of the buyers, the SGP (Scam Gaming Profit) can not be reduced below zero so that gaming remains profitable even with the highest PLR possible.

PLR	OMU	LTS	BSL	SGP
0	0.92	0.10	0.14	2.01
2	0.95	0.06	0.10	1.39
5	0.97	0.04	0.08	1.21
10	0.97	0.04	0.09	1.27
15	0.97	0.04	0.08	1.14
30	0.97	0.03	0.08	1.11
70	0.98	0.03	0.07	1.07
100	0.98	0.03	0.07	1.05

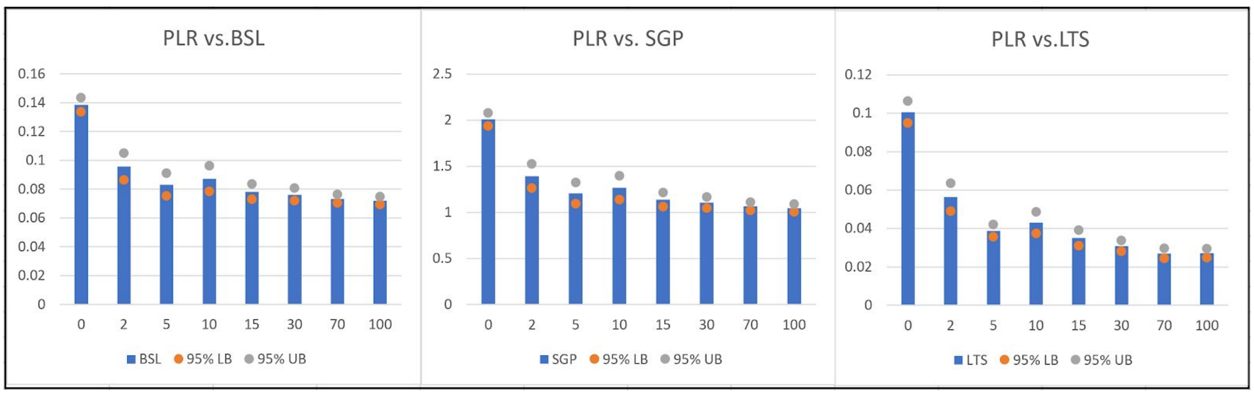


Table and charts for dependency of financial metrics on probability for leaving ratings for “Regular” reputation system using adaptive-simulation. The charts show 95% confidence intervals.

The tabular data and charts below render performance of the financial metrics. We made multiple simulations and compared the effects that various parameter settings have upon performance metrics. We then compared results between “Regular” and “Weighted” reputation system, TOM/SOM (time/spendings on the market) based ones, “Anti-biased”, “Predictive” and “Vendor Impact” reputation system. Optimisation was targeted at making OMU (Organic Market Utility) higher and the other metrics lower. Our simulation results show that buyers were satisfied with the products that they bought at the price that they bought them, as one can tell from the BSL column.

Reputation System Type	OMU	LTS	BSL	SGP
None	0.99	0.01	0.06	0.83
Regular	0.97	0.03	0.08	1.11
Weighted	0.99	0.01	0.03	0.37
TOM	0.99	0.01	0.03	0.36
SOM	0.99	0.02	0.03	0.46
Anti-biased	1.00	0.00	0.02	0.25
Predictive	0.99	0.01	0.03	0.40
Vendor Impact	0.99	0.01	0.03	0.37

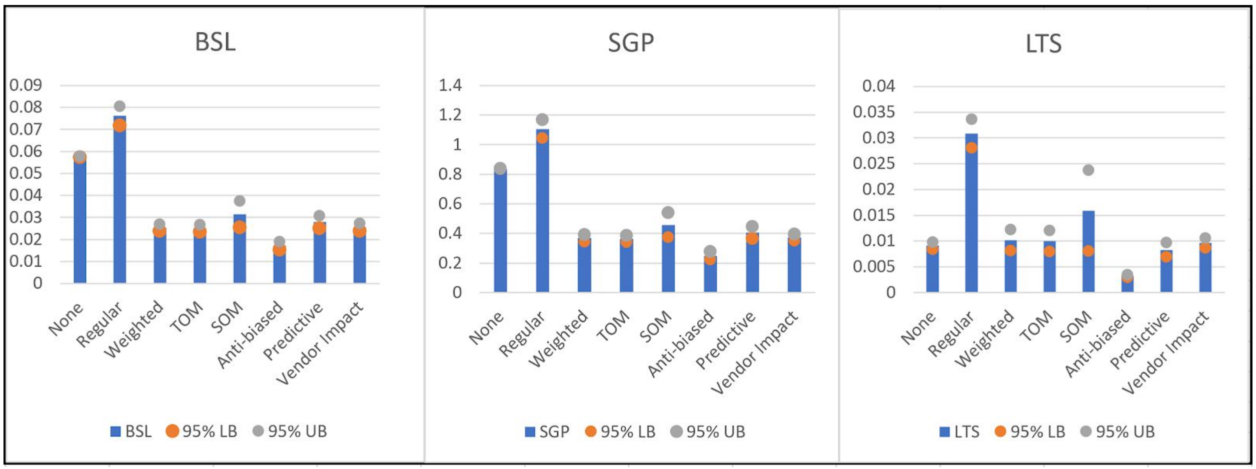


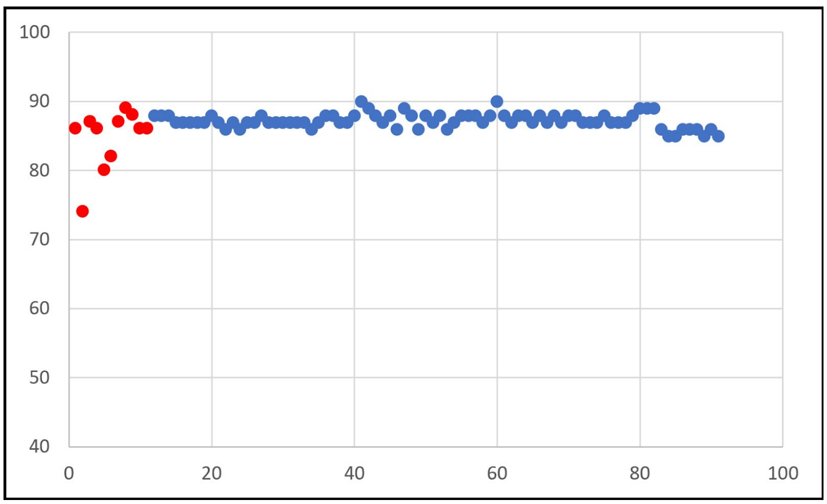
Table and charts presenting performance of financial metrics for different reputation systems using adaptive simulation. The charts show a 95% confidence interval for the highest and lowest the true values could be (had we repeated the simulations indefinitely).

The tabular data and charts above show that use of the “Regular” reputation system makes all financial metrics worse than in the case in which no reputation system is used. This result is caused by reputation gaming redirecting the market to dishonest providers, thereby increasing their profits (SGP), decreasing the volume of honest market (OMU), and causing losses for buyers (LTS and BSL). We also see that most reputation system configurations, including “Anti-biased”, Weighted, TOM, “Predictive”, and “Vendor Impact” improve the financial metrics. The LTS column shows that the best “Anti-biased” reputation system configuration reduced the total market volume spent on scams to zero and made the OMU approach 1.00. We see that while a regular reputation system results in a profit for dishonest suppliers in the SGP, the best reputation systems greatly reduce their profits.

Comparing BSL, SGP and SGL, we note that the regular reputation system, as well as weighted and SOM, underperform compared with anti-biased and predictive reputation systems. If we choose optimal settings however, SOM based system proves promising under certain conditions. We find similar results when considering OMU and LTS.

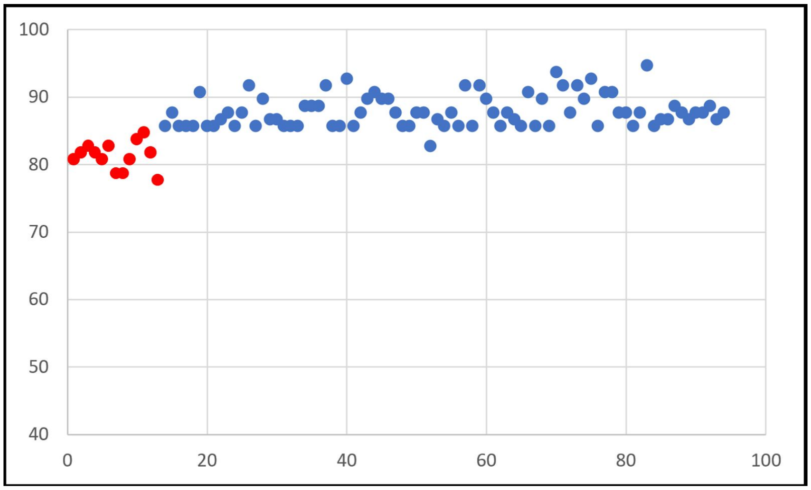
The following charts display reputation ranks of all sellers present on the market arranged so that scamming sellers are on the left (red dots) and honest sellers are on the right (blue dots). The seller reputation ranks in this case are computed as averages across reputation ranks of all products listed under respective sellers. The scamming sellers sell very low quality products, but attempt to boost their scores gaming the reputation system “sponsoring” some of the buyers to write reviews. The blue sellers offer quality goods and don’t need to game the reputation system. These charts, which include only active suppliers, present normalized ranks. The relative ranks of the sellers in the reputation system are what is important in a given study, because sellers that are higher on the chart are more recommended by the reputation system than sellers that are lower on the chart.

The first chart below refers to what we call the “Regular” reputation system. Note that many of the dishonest sellers are ranked higher by this standard reputation system than many of the honest sellers.



Distribution of reputation ranks for dishonest (red, left) and honest (blue, right) sellers when “Regular” reputation system is in use.

The use of an “Anti-Biased” reputation system, taken as an example, turns to be a game changer, as the following chart shows. Note that the dishonest sellers are ranked lower than the honest ones, as they should be. This reputation system is able to expose dishonest suppliers despite their purchasing reviews, switching their products and changing their identities during the simulation.



Distribution of reputation ranks for dishonest (red, left) and honest (blue, right) sellers when “Anti-biased” reputation system is in use.

Qualitatively similar results above were confirmed with both non-adaptive and adaptive simulations. In particular, the introduction of the “Regular” reputation system makes all metrics worse compared to the case when there is no reputation system at all while using the “Anti-biased” reputation system configuration always improves all of the metrics.

The following parameters provided the best results for our adaptive simulations in the “Anti-biased” setup.

- Denomination = TRUE; this is the parameter which normalizes new (updating) ranks if set to true.
- Decayed = 0.1; Ranks decay at a rate of 10% per period (1 period means one day in our simulation).
- Default = 0.5 for “Anti-biased”; Default ranks of each product are at 0.5 and then change based on conservatism and feedback.
- Conservatism = 0.5. Conservatism identifies how much weight in rank calculation we give to the older ranks and how much the new ratings affect the updated ranks.

Conclusions

We designed and evaluated a reputation system based on the **liquid rank algorithm**, which aims at providing accurate recommendations for products on online marketplaces. This reputation system is particularly well equipped to tackle online reputation manipulation, which frequently occurs in modern online marketplaces. Besides the liquid rank algorithm features we previously presented, we added new ones including anti-biasedness of buyers, and predictiveness. We also created new, improved simulations based on real market conditions derived from our detailed market research. The simulations let us conclude the following.

- The probability of leaving a rating greatly impacts the performance of our reputation system. When more real buyers provide feedback on their experiences we can better determine product quality, and makes it easier to weed out dishonest buyers from honest ones. As a result of our simulations, we strongly recommend that online marketplaces encourage buyers to leave ratings as often as possible.
- We also recommend using “Anti-biased” reputation system configuration, with the reported parameter settings, on marketplaces with ratings design and behavior similar to those of Amazon, as this proved to be the best at improving financial metrics.

Based on our research, more studies are required in the following directions.

- The need of further studies combining configurations of reputation system types and fine tuning of the parameters, so that better resistance to the reputation cheating is found.
- There is a need to design and implement another system layer to control overall performance and adjust the setup as market conditions change.
- In the reported studies it was assumed that honest vendors serving high-quality goods never cheated. However, it is known that in the real world even honest vendors must game their ratings in order to stay in the gamified market environment. We need to simulate this sort of market in order to create a reputation system that allows honest sellers to engage in as little gaming themselves and minimizing the SGL (Scam Gaming Loss) metric.