Reciprocity-induced bias in digital reputation

Giacomo Livan,[1,2] Fabio Caccioli,[1,2] and Tomaso Aste[1,2]

[1]*Department of Computer Science, University College London, Gower Street, London WC1E 6EA, UK*
[2]*Systemic Risk Centre, London School of Economics and Political Sciences, Houghton Street, London WC2A 2AE, UK*

(Dated: June 9, 2016)

The peer-to-peer (P2P) economy relies on establishing trust in distributed networked systems, where the reliability of a user is assessed through digital peer-review processes that aggregate ratings into reputation scores. Here we present evidence of a network effect which biases the digital reputations of the users of P2P networks, showing that P2P networks display exceedingly high levels of reciprocity. In fact, these are so large that they are close to the highest levels structurally compatible with the networks' reputation landscape. This shows that the crowdsourcing process underpinning digital reputation is significantly distorted by the attempt of users to mutually boost reputation, or to retaliate, through the exchange of ratings. We show that the least active users are predominantly responsible for such reciprocity-induced bias, and that this fact can be exploited to suppress the bias itself.

I. INTRODUCTION

The digital economy is increasingly self-organizing into a "platform society" [1, 2] where individuals exchange knowledge, goods, and resources on a peer-to-peer (P2P) basis. Indeed, in recent years a number of well-established business-to-consumer sectors, such as the taxi and hotel industries [3], have been disrupted by the emergence of the novel sharing economy P2P marketplaces.

P2P platforms rely on trust between their users. Trust is typically established by developing a reputation through digital peer-review mechanisms that allow users to rate their peers and their activity. Given the spectacular growth of the P2P paradigm in recent years and its expected growth in the near future, digital reputation will increasingly become central in our online lives, as it will grant or prevent access to substantial economic opportunities. Hence, it is crucial to ensure the fairness of digital peer-review systems [4], and to ensure that individual reputation scores reflect the performance of users in an accurate and unbiased way.

Being decentralized, P2P systems are often thought to promote more economic freedom and more democratization. Yet, their current lack of regulation exposes them to a number of biases which can distort their functioning [5–8]. Game theoretic considerations [9, 10], and plenty of anecdotal evidence, suggest that users are often incentivized to reciprocate ratings, i.e. to exchange positive ratings to mutually boost their reputation and to retaliate after receiving negative ones. For instance, the "5 for 5" practice of Uber drivers and passengers, i.e. agreeing on exchanging 5 star ratings at the end of a ride, is a common firsthand experience of such a practice [11]. Similar phenomena are also well documented empirically in the interactions between buyers and sellers in online marketplaces such as eBay [12, 13].

Reputation in online marketplaces is known to affect a buyer's willingness to pay for goods or services [13]. In this respect, reciprocity-induced distortions effectively introduce *externalities* in P2P platforms, as they prevent users from making informed *ex ante* decisions about their peers. Moreover, the anticipation of retaliatory behavior can discourage users from posting negative reviews. As a matter of fact, it is well known that online ratings are often skewed towards positive values [7, 14]. All in all, reciprocity may deteriorate the overall information content in P2P platforms, and therefore poses a threat to their fairness and transparency.

The goal of the present paper is to tackle the above issue by putting forward a quantitative method to discern the information content of ratings in P2P platforms. We do so by taking a network perspective in order to quantify the extent of reciprocity-induced biases in online reputation systems. We investigate three case studies of P2P platforms with binary interactions (i.e. users rate their peers either positively or negatively) that can be conveniently represented in terms of signed networks: we represent a platform user as a node in the network, and we represent an endorsement (dislike) of node j's activity from node i as a directed link $i \to j$ carrying a $+1$ (-1) weight. We consider the following platforms: **(i) Slashdot**, a technology news website whose users label each other as "friend" or "foe" based on their public comments; **(ii) Epinions**, a platform for crowdsourced consumer reviews; **(iii) Wikipedia**, where positive and negative edits are interpreted as signed links. Such systems have attracted considerable attention [15–17], as they offer a natural laboratory to test social theories for systems with antagonistic interactions, such as social balance theory [18–20] and consensus formation [21].

In order to suppress noisy contributions from casual platform users, we restrict the networks to a high participation core of users who have both given and received at least ten ratings within the core (see the Appendix for more information about the statistical properties of such core).

II. REPUTATION AND RECIPROCITY

We say that a positive (negative) link $i \to j$ is reciprocated if a positive (negative) link $j \to i$ also exists. We then define positive (negative) reciprocity ρ^+ (ρ^-) as the fraction of positive (negative) links that are reciprocated. This definition is meaningful in sparse networks such as the ones we will work with. Let us remark, however, that it would be less informative in dense networks, where reciprocated links exist due to structural constraints (i.e. the number of links is so large that a substantial fraction must be reciprocated), and could be replaced by the notion of reciprocity put forward in [22], which discounts density-related effects. In our case, the two definitions coincide for all practical purposes.

The ratings received by a node i can be aggregated into a measure of reputation R_i, which we define as the difference between the numbers of positive and negative ratings received by node i, divided by its total number of received ratings (see Eq. (A3) in the Appendix). By construction, we have that $R_i \in [-1, 1]$, where $R_i = 1$ ($R_i = -1$) for a node that has received positive (negative) ratings only.

III. EXCESS RECIPROCITY

In order to quantify excess reciprocity we compare the reciprocity observed in the empirical networks with a "basal" level $\rho_0^\pm$ of reciprocity compatible with the overall reputation landscape of the network. We achieve this by reshuffling links in the network while preserving the numbers of positive and negative ratings received and given by each individual node (see Fig. 1). This procedure is reminiscent of the directed configuration model from the literature on complex networks [23], and amounts to randomly redirecting ratings, hence destroying correlations between raters and ratees, while preserving both the reputation of each node and the system's heterogeneity in terms of user activity (i.e. number of positive/negative ratings received and given, see Fig. 4 in the Appendix). In other words, $\rho_0^\pm$ quantify the minimal reciprocity that needs to be in the system due to its density and the constraints on each node's reputation. In the Appendix we discuss a broader class of null models. The reshuffling procedure described here corresponds to the case $\beta = 0$ of the general null model class we consider in the Appendix.

TABLE I: Comparison between the positive and negative reciprocity $\rho^\pm$ observed in the three networks we analyze and the 99% confidence level intervals for the corresponding "basal" levels $\rho_0^\pm$ and saturation levels $\rho_{\rm SAT}^\pm$ obtained under a null hypothesis of random link rewiring constrained to preserve each user's reputation.

	ρ^+	ρ^-	ρ_0^+	ρ_0^-	$\rho_{\rm SAT}^+$	$\rho_{\rm SAT}^-$
Slashdot	41.3%	15.9%	[2.28; 2.53]%	[0.35; 0.62]%	[46.6, 47.0]%	[25.6; 26.2]%
Epinions	42.4%	7.70%	[2.33; 2.49]%	[1.08; 1.42]%	[48.2; 48.4]%	[24.9; 25.3]%
Wikipedia	17.6%	8.50%	[3.01; 3.27]%	[1.84; 2.45]%	[36.8; 37.2]%	[48.4; 49.1]%

Table I shows both the positive and negative reciprocity $\rho^\pm$ we measure in the original networks (columns 1 and 2), and the corresponding basal levels $\rho_0^\pm$ (columns 3 and 4). With respect to the null benchmarks, we observe very large values of both positive and negative reciprocity, well outside the 99% confidence level intervals of the basal reciprocity levels in the null model. The significantly higher values of positive reciprocity and the overall relative abundance of positive ratings (see Table III in the Appendix) motivate the greater attention we will devote to positive reciprocity in the following.

IV. RECIPROCITY SATURATION

After estimating the basal levels of reciprocity, we now turn our attention to the opposite issue. Namely, could P2P systems accommodate more reciprocity than we actually observe? We answer this question by resorting again to sampling from a suitable ensemble of null models. As before, we perform a random rewiring of positive and negative links while preserving every node's reputation. Additionally, we now require the systems to reach a predefined

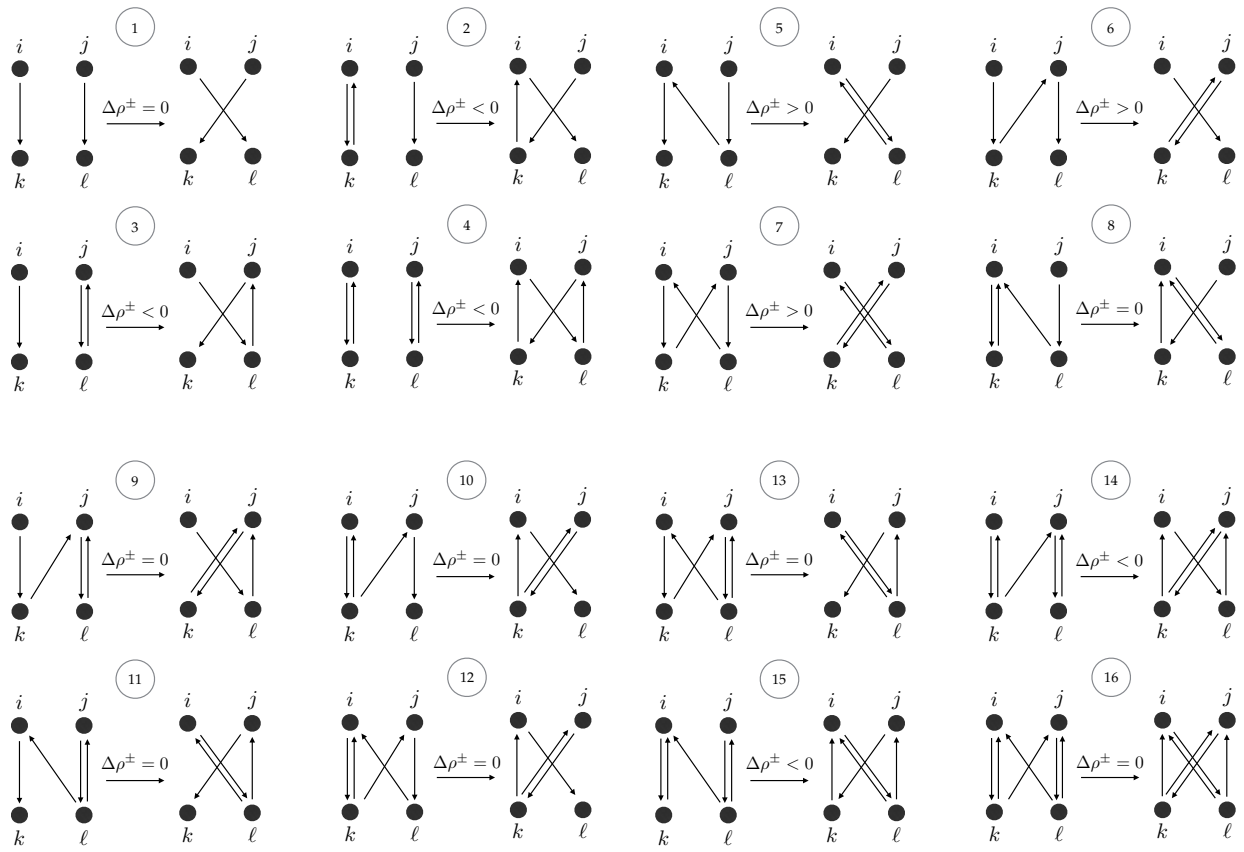

FIG. 1: **List of possible moves allowed by our rewiring procedure.** In each case (i, k) and (j, ℓ) are two pairs of randomly selected nodes connected by two directed links $i \to k$ and $j \to \ell$ of the same sign. Such links are rewired from k to ℓ and from ℓ to k, respectively (provided that links $i \to \ell$ and $j \to k$ do not already exist) with a probability that depends on the change in positive/negative reciprocity $\Delta\rho^{\pm}$ that the particular move would entail, on a reciprocity target $\tau^{\pm}$, and on an "intensity of choice" parameter $\beta \geq 0$ (see Appendix for more details). When $\beta = 0$ such a probability is equal to 1/2 regardless of the target, i.e. links are rewired at random. When $\beta \to \infty$ the only accepted rewiring operations are those that move the system's reciprocity towards the target. Depending on the particular draw of nodes, other links connecting the two pairs (i.e. the links $\ell \to i$ and/or $k \to j$) or reciprocated links (i.e. $k \to i$ and/or $\ell \to j$) might exist. Therefore, a particular rewiring move might increase reciprocity (moves 5, 6, 7), decrease reciprocity (moves 2, 3, 4, 14, 15) or leave reciprocity untouched (moves 1, 8, 9, 10, 11, 12, 13, 16).

reciprocity target $\tau^{\pm 1}$.

By increasing the reciprocity target above the original networks' reciprocity levels, we find that all three platforms reach a saturation both in positive and negative reciprocity, i.e. the networks run out of links that can be used to reciprocate while preserving each node's reputation. In Table I (columns 4 and 5) we report such values, which we denote as $\rho^{\pm}_{\rm SAT}$.

As reported in Table I we find that both Slashdot and Epinions run out of positive links to create additional reciprocity shortly after the target τ^{+} exceeds the actual positive reciprocity ρ^{+} (i.e. the ratio $\rho^{+}_{\rm SAT}/\rho^{+}$ is roughly equal to 1.15). On the contrary, Wikipedia could accommodate values of reciprocity much larger than its own ρ^{+}, i.e. the ratio $\rho^{+}_{\rm SAT}/\rho^{+}$ is roughly equal to 2.10, which can be related to the different nature of the interactions. Indeed, interactions in Slashdot, where positive and negative links correspond to users tagging each other as "friend"

[1] This corresponds to the case of very large β (namely, $\beta L^{+} = 10^{10}$, where L^{+} denotes the total number of positive links) in the null models outlined in the Appendix. From a practical point of view, this amounts to implementing a soft constraint, i.e. requiring the system to produce an average level of reciprocity equal to $\tau^{\pm}$ with some small fluctuations around it.

or "foe", encourage backscratching and retaliatory behavior, whereas a collaborative environment such as Wikipedia is capable of sustaining more than twice the positive reciprocity it displays in real life. This picture is corroborated by the findings on negative reciprocity as well, where the ratio $\rho^-_{\rm SAT}/\rho^-$ increases from roughly 1.5 to almost 6 as progressing from Slashdot to Wikipedia. The general remark one can make from such values is that more polarized P2P environments (such as Slashdot) exist closer to their reciprocity saturation levels.

V. PRODUCTION OF REPUTATION THROUGH RECIPROCITY

In the previous sections we have provided evidence that P2P systems display excessive reciprocity, and, in fact, could hardly sustain higher reciprocity levels. We now ask whether reciprocity biases reputation in P2P systems, and, if so, to what extent. To this end, let us divide the links in the networks we study into the following four categories: unreciprocated positive links, unreciprocated negative links, reciprocated positive links, and reciprocated negative links. We denote the number of links belonging to each category as, respectively, Φ^+, Φ^-, Γ^+, and Γ^-.

Unreciprocated links correspond to the part of reputation built on ratings received from peers that did not receive a rating in return. As such, they can be reasonably assumed to represent fair and objective peer assessments, and their contribution to reputation can be thought of as a proxy of a user's "true" reputation. On the other hand, reciprocated links could be due to backscratching and retaliation, and could introduce distortions in user reputation. We are precisely interested in estimating the entity of such distortions.

In order to disentangle the different contributions, for each link category we focus on the average contribution to reputation coming from one link belonging to it. Let us introduce the total reputation in the network $R = \sum_{i=1}^{N} R_i$ and write

$$R = \Phi^+\lambda^+_\Phi - \Phi^-\lambda^-_\Phi + \Gamma^+\lambda^+_\Gamma - \Gamma^-\lambda^-_\Gamma \ , \tag{1}$$

where $\lambda^\pm_\Phi$ ($\lambda^\pm_\Gamma$) are the average contributions to reputation associated with the existence of a positive/negative unreciprocated (reciprocated) link.

In Table II we report the values of the above quantities measured in the three networks we analyze. As it can be seen, in all cases we have $\lambda^+_\Gamma > \lambda^+_\Phi$, i.e. on average reciprocated positive links contribute more to the formation of user reputation than unreciprocated ones. We instead find mixed signatures in the case of negative activity: only in Slashdot, where negative interactions are genuinely hostile (users labeling their peers as "foes"), we observe $\lambda^-_\Gamma > \lambda^-_\Phi$, i.e. that reciprocated negative links play a larger role in shaping reputation than unreciprocated ones. We interpret this as a signature of retaliatory behavior.

TABLE II: Average contribution to reputation from each link category, as defined in Eq. 1: $\lambda^\pm_\Phi$ denote the average contribution from a positive/negative unreciprocated link, while $\lambda^\pm_\Gamma$ denote the average contribution from a positive/negative reciprocated link. All values in the Table have been multiplied by 10^2.

	λ^+_Φ	λ^+_Γ	λ^-_Φ	λ^-_Γ
Slashdot	3.23	3.55	3.40	5.22
Epinions	1.56	2.20	2.36	1.57
Wikipedia	2.70	3.25	3.62	3.21

We test the statistical significance of the above findings by resorting again to null hypotheses of random link rewiring. We use the same null models used to produce the data pertaining to reciprocity saturation in Table I, spanning the whole possible range of positive reciprocity targets τ^+, from zero to saturation. Fig. 2 shows the average behavior of the contributions to reputation coming from positive reciprocated and unreciprocated links. A number of relevant results can be deduced from this Figure.

First, the behavior of the two quantities as functions of τ^+ is radically different: the contribution from reciprocated activity λ^+_Γ is monotonically increasing, as one would intuitively expect, whereas the contribution from unreciprocated activity λ^+_Φ displays a non-monotonic behavior and attains a maximum in correspondence of a certain reciprocity target (which depends on the particular network).

Second, throughout the whole reciprocity range, we find that in the empirical networks the contribution to reputation from unreciprocated activity is under-expressed with respect to our null hypothesis. Symmetrically, the contribution from reciprocated links is systematically over-expressed. Both these conditions hold under a much more general class of null hypotheses (see the Appendix). This highlights the existence of a *reciprocity bias*, i.e. that reciprocated activity plays an exceedingly large role in shaping reputation at the aggregate level, which can quantified by the differences between the rates λ^+_Φ and λ^+_Γ as measured in the actual networks and in null models.

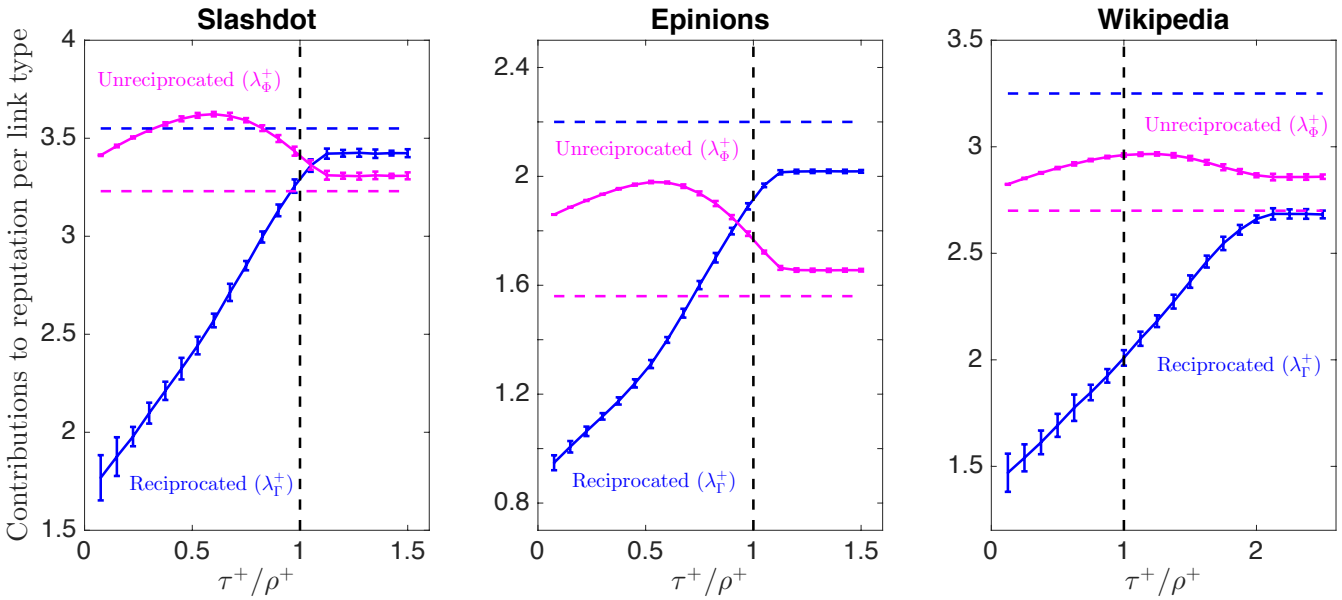

FIG. 2: **Demonstration that reputation is affected by a reciprocity bias**. Solid lines show the average contribution to reputation from unreciprocated positive ratings (λ_Φ^+, pink) and reciprocated positive ratings (λ_Γ^+, blue) obtained under a null hypothesis of random link rewiring constrained to preserve each user's reputation and to produce a predefined positive reciprocity target τ^+. The behavior of λ_Φ^+ and λ_Γ^+ is shown as a function of the ratio between the reciprocity target τ^+ and the positive reciprocity ρ^+ measured in the actual platforms (column 1 of Table I). Error bars correspond to 99% confidence level intervals. Dashed lines correspond to the values of λ_Φ^+ (pink) and λ_Γ^+ (blue) measured in the actual platforms (i.e. to the values reported in columns 1 and 2, respectively, of Table II). All values in the plots have been multiplied by 10^2. The fact that the contribution from reciprocated (unreciprocated) activity in the actual platforms is systematically lower (higher) than under our null hypothesis highlights the existence of the reciprocity-induced bias.

Third, in null models the relative importance between reciprocated and unreciprocated activity is overturned with respect to the one observed in the actual networks. As shown in Table II, one positive reciprocated link always contributes more to reputation, on average, than an unreciprocated one (i.e. $\lambda_\Gamma^+ > \lambda_\Phi^+$). In contrast, under our null hypothesis the opposite holds over a wide range of the reciprocity target τ^+. Namely, one has $\lambda_\Gamma^+ < \lambda_\Phi^+$ almost up to the saturation threshold in Slashdot and Epinions, whereas for Wikipedia this is the case for any reciprocity target. Notably, both in Slashdot and Wikipedia one has $\lambda_\Gamma^+ < \lambda_\Phi^+$ even when the reciprocity target is kept equal to its value in the actual networks, i.e. when $\tau^+ = \rho^+$ (dashed vertical lines in Fig. 2). This is a case where our null hypothesis entails the injection of a minimal amount of randomness into the system, as the only rewiring operations allowed are those that do not change reciprocity even at the local level (see Fig. 1 and its caption for more details). Yet, these operations are sufficient to overturn the relative importance between reciprocated and unreciprocated activities. This makes it clear that real-life rating dynamics drive the system towards very peculiar states, whose fragility we quantify in the next Section.

VI. SUPPRESSING THE RECIPROCITY BIAS THROUGH RANDOM LINK ELIMINATION

In the previous section we explored the production of reputation in P2P environments at the macroscopic level. In the three networks we have studied we have consistently observed the presence of a reciprocity bias which makes the contribution to reputation from positive reciprocated activity prevalent with respect to that coming from unreciprocated activity (i.e., $\lambda_\Gamma^+ > \lambda_\Phi^+$ in our notation). Yet, our analysis of null models shows that such a feature disappears as soon as the networks are slightly randomized. In other words, this suggests that P2P dynamics drive the networks towards very "atypical" states whose main features are not robust to small perturbations.

Fig. 3 shows the behavior of the average contribution to reputation from reciprocated (λ_Γ^+) and unreciprocated (λ_Φ^+) positive links upon the removal of small fractions of reciprocated positive links. Notably, the deletion of 3% of the reciprocated positive links (i.e. slightly more than 1.2% of the total positive links) in Slashdot is enough to make the contributions to reputation from reciprocated and unreciprocated links statistically compatible. The same result is achieved by removing roughly 8% of the reciprocated positive links in Epinions, and roughly 11% of the

reciprocated positive links in Wikipedia (corresponding, respectively, to 3.3% and 1.9% of the overall positive links). Furthermore, one can also see from Fig. 3 that statistical compatibility between λ_Γ^+ and λ_Φ^+ as measured in the full networks and in the networks after the removal of a few links is lost extremely fast, i.e. by removing less than 1% of the reciprocated positive links.

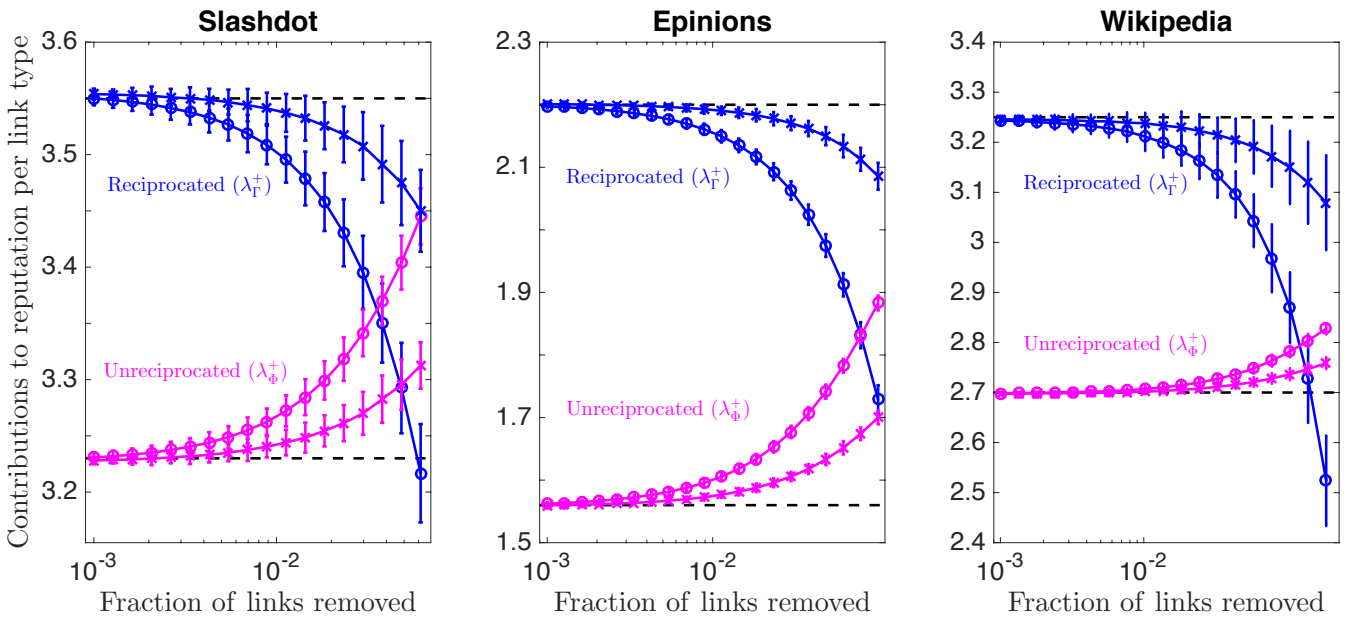

FIG. 3: **Demonstration that the elimination of a small fraction of ratings suppresses the reciprocity bias**. Solid lines show the average contribution to reputation from unreciprocated (λ_Φ^+, pink) and reciprocated (λ_Γ^+, blue) positive links as a function of the fraction of reciprocated positive links removed from the network. Circles represent the behavior of such quantities when a random node selection protocol is followed, i.e. nodes are chosen at random with uniform probability and reciprocated positive links between them, if any, are removed. Crosses refer instead to a random link selection protocol, where links are removed with uniform probability. In the former case the majority of links removed are between low degree nodes, whereas in the latter case the elimination procedure targets hubs with higher probability. The dashed lines represent the values of λ_Γ^+ (upper line) and λ_Φ^+ (lower line) in the original networks.

Let us remark that the protocol chosen to carry out the random link elimination procedure is crucially important to keep the fraction of removed links so low. Indeed, an efficient protocol consists in selecting *pairs of nodes* at random, checking whether a reciprocated link between them exists, and, if so, removing it. Eliminating reciprocated links at random with uniform probability requires instead the elimination of a much larger number of links, as it can be seen in Fig. 3. Given the heavy tailed nature of the distributions of ratings given and received by each node (see Fig. 4 in the Appendix), this means that the most efficient way to decrease the overall contribution to reputation from reciprocated activity is to eliminate the reciprocated links between nodes with low degrees. In contrast, removing any link with equal probability amounts to preferentially removing links between hubs in the network, i.e. high activity users.

The above result might seem unfair at first, as on average it penalizes users with a low number of ratings, whereas it leaves high activity users with many reciprocated links relatively untouched. However, let us remark that the networks we analyze represent a high participation core of nodes with at least ten incoming and outgoing links, i.e. the contribution from casual platform users have been filtered out. Moreover, such a protocol makes sense from the viewpoint of user incentives. Indeed, a newcomer to a platform is more incentivized to reciprocate in order to boost visibility in the network, whereas a high activity user with a good reputation has a little marginal gain from an additional rating. In this respect, removing ratings given/received by low activity users is key to suppressing the reciprocity bias and improving the average quality of ratings.

VII. DISCUSSION

The present paper provides a first systematic study of the network effects shaping digital reputation in P2P platforms. In this work we have tested the statistical significance of a number of empirical facts consistently observed in

the three platforms we studied. We have done so by investigating a range of null models designed to preserve the number of positive/negative ratings given and received by each user, hence each user's individual reputation, while probing different rating patterns that could have produced them. This effectively amounts to exploring "alternate realities" of P2P systems, while still keeping their heterogeneity fully intact at the level of individual users.

The overarching question we addressed in this framework is whether P2P platform users excessively engage in rating reciprocity in order to improve their reputation or to affect that of others. We do find that reciprocity, especially in the positive case, is substantially over-expressed with respect to a null benchmark. Moreover, in all three systems we find that reciprocated ratings contribute more to reputation than unreciprocated ones. This is at odds with what we observe in a vast class of null hypotheses that preserve individual reputations, where we always find that unreciprocated activity dominates the production of reputation. In other words, this shows that the same individual reputations are compatible with very different rating patterns between the users. In conclusion, the local structure of the empirical networks are responsible for the distortions observed at the macroscopic level.

The above point suggests that P2P systems exist in very peculiar states. Indeed, the contribution to reputation from reciprocated activity is systematically over-expressed with respect to all of the null hypotheses we consider (see Fig. 5 and Fig. 6 in the Appendix), and a small random perturbation is enough to make unreciprocated activity the prevalent contribution. This point is evocative of other results concerning the beneficial effects of randomness in complex systems (see, e.g., [24] or [25] and references therein), which suggests that an effective policy to prevent users from building reputation through excessive reciprocity in our simplified framework would be that of injecting a small quantity of randomness into the system. We validated this hypothesis by carrying out a random link elimination procedure in the three networks we analyzed, which shows that the removal of reciprocated links between users with a low number of ratings, hence highly incentivized to boost reputation, is most effective.

Our investigation highlights that interactions of a different nature (i.e., collaborative vs antagonistic) lead to different network signatures. Both in Slashdot and Epinions, suppressing reciprocity unquestionably makes unreciprocated ratings the prevalent contribution to reputation, up to a point (roughly corresponding to 50% of the reciprocity observed in the actual networks, see Fig. 2) where the contribution from unreciprocated activity reaches a maximum. Conversely, and rather paradoxically, in the Wikipedia network reciprocity has to be increased with respect to its original level in order to reach the maximum contribution from unreciprocated activity. Indeed, Wikipedia reaches the highest average contribution to reputation from unreciprocated activity for reciprocity values higher than the one observed in the actual network. We speculate that this is due to the different outcomes that such networks aim to achieve. In essence, Wikipedia is a collaborative *content-driven* environment whose users cooperate to the creation of a common good, i.e. knowledge. In contrast, Epinions and Slashdot have a more personal trait, as in both cases interactions are *opinion-driven*: users form relationships based on the endorsement or rejection of their peers' views. Our results suggest that an increase in reciprocity in a content-driven environment might lead to increased collaboration and, ultimately, to an improved quality of the ratings exchanged by the users. This aspect certainly deserves further attention through the analysis of other P2P networks.

The systems we study in this paper are simpler than the most popular P2P platforms where users build a peer-review based reputation, such as Uber and Airbnb. Yet, they retain the full complexity of those richer environments, both in terms of interaction patterns and user heterogeneity. It is precisely because of such a "stylized-yet-complex" nature that we chose signed networks as templates for P2P systems.

All in all, our analyses show that P2P systems are plagued by biases, and far from the ideal of transparency they are often thought to represent. We have shown that the most widely adopted reputation metrics, i.e. those based on naive rating aggregation, are particularly vulnerable to distortions, which we have related with the presence of non-trivial network effects and motifs. In this respect, our work highlights the yet largely untapped potential that network science applications have in the digital economy domain, and suggests that novel, network-based, notions of reputation could be the way to ensure the fairness that P2P systems promise to deliver.

Acknowledgments

We acknowledge support from the Economic and Social Research Council (ESRC) in funding the Systemic Risk Centre (ES/K002309/1). Giacomo Livan acknowledges support from an EPSRC Early Career Fellowship in Digital Economy (Grant No. EP/N006062/1).

Appendix A: Reputation and reciprocity

We describe a signed network of N nodes by means of a square $N \times N$ adjacency matrix A, whose entries are such that $A_{ij} = 1$ ($A_{ij} = -1$) if node i has given a positive (negative) rating to node j, and $A_{ij} = 0$ otherwise. With this

notation, one can then define the quantities [2]

$$\phi_i^{\pm} = \sum_{j=1}^{N} \Theta(\pm A_{ji}) \left[1 - \Theta(\pm A_{ij})\right] \tag{A1}$$

$$\gamma_i^{\pm} = \sum_{j=1}^{N} \Theta(\pm A_{ji}) \Theta(\pm A_{ij}) \ ,$$

which represent, respectively, the number of incoming unreciprocated links and the number of incoming reciprocated links of a node i. We then define $\Phi^{\pm} = \sum_{i=1}^{N} \phi_i^{\pm}$ and $\Gamma^{\pm} = \sum_{i=1}^{N} \gamma_i^{\pm}$ as the total number of links belonging to each category.

Given our focus on reputation we filter the three network datasets in order to suppress noisy contribution from casual platform users. To this end we restrict the networks to a high participation core, which we obtain by iteratively removing nodes with less than ten outgoing and incoming links (we checked that all our results are consistent across different thresholds for the determination of the core). Table III reports, for each network, the number of nodes, the overall number of positive links ($L^+ = \Phi^+ + \Gamma^+$), and the overall number of negative links ($L^- = \Phi^- + \Gamma^-$) we end up with after such filtering[3]. As it can be seen, in all three networks positive ratings represent roughly $80 - 90\%$ of the total number of links.

TABLE III: Number of users N and number of positive (L^+) and negative (L^-) ratings in the high participation core of the three platforms we analyze.

	N	L^+	L^-
Slashdot	4611	105115	29190
Epinions	8732	425377	40996
Wikipedia	5538	160606	29327

As it is often the case in networked systems, links are not distributed evenly between the nodes of the platforms we analyze. Namely, both the distribution of the ratings received and the distribution of the ratings given by each node[4] are heavy tailed, and reasonably well fitted by power laws of the form $p(x) \sim x^{-\alpha}$ (see Fig. 4).

---

[2] $\Theta(x)$ denotes the step function such that $\Theta(x) = 1$ for $x > 0$ and $\Theta(x) = 0$ otherwise.

[3] The full empirical networks can be downloaded from http://konect.uni-koblenz.de/

[4] The number of ratings received and given by node i are given, respectively, by the sums $\sum_{j=1}^{N} |A_{ji}|$ and $\sum_{j=1}^{N} |A_{ij}|$

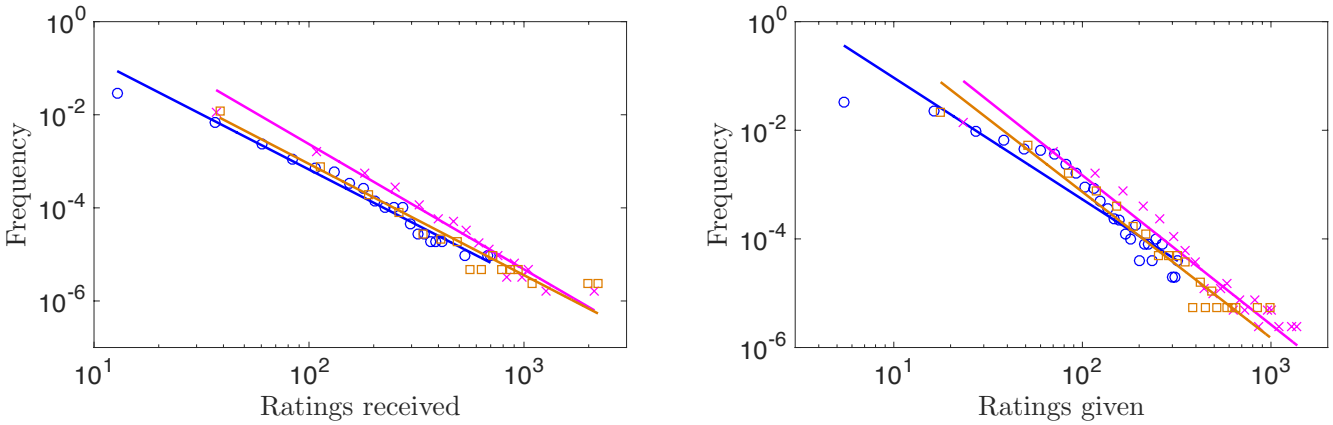

FIG. 4: **Power law distribution of the ratings received and given by each user**. Empirical distribution of the ratings received (left panel) and given (right panel) by each node. Circles represent Slashdot data, crosses represent Epinions, and squares represent Wikipedia. In each case the empirical distribution is reasonably well fitted by a power law of the form $p(x) \sim x^{-\alpha}$ (shown as solid lines with colors matching those of the empirical distributions). For the distribution of received ratings we find $\alpha = 2.37$ (Slashdot), $\alpha = 2.69$ (Epinions), and $\alpha = 2.39$ (Wikipedia). In the case of given ratings we find $\alpha = 2.24$ (Slashdot), $\alpha = 2.75$ (Epinions), and $\alpha = 2.69$ (Wikipedia).

We denote positive (negative) reciprocity as ρ^+ (ρ^-), and we define it as the relative frequency of links $i \to j$ that have a matching link $j \to i$ of the same sign. Using the above definitions we have

$$\rho^{\pm} = \frac{\Gamma^{\pm}}{L^{\pm}} \ . \tag{A2}$$

Using the same quantities introduced above, we define the reputation of node i as

$$R_i = \frac{\phi_i^+ - \phi_i^- + \gamma_i^+ - \gamma_i^-}{\phi_i^+ + \phi_i^- + \gamma_i^+ + \gamma_i^-} \ , \tag{A3}$$

i.e. as the difference between the number of positive and negative ratings received (both reciprocated and unreciprocated) normalized by the overall number of ratings received. The above definition of reputation is such that $-1 \leq R_i \leq 1$, where $R_i = 1$ ($R_i = -1$) for a node that has received positive (negative) ratings only.

Appendix B: Null models

In order to establish the significance of the quantities we measure, we define ensembles of null network models that depend on two parameters. Namely, we define a positive/negative target reciprocity $\tau^{\pm}$, and we introduce a cost function $H(\tau^{\pm}) = [L^{\pm}(\rho^{\pm} - \tau^{\pm})]^2$ to measure the distance between the current positive/negative reciprocity in the network and the target. Starting from the empirical networks, we perform rewiring operations in order to decrease the cost function's value, i.e. to make the networks' reciprocity converge to the predefined target. We do so in a probabilistic manner: we iteratively propose random rewiring operations and we accept them with probability

$$p(\beta, \tau^{\pm}) = \frac{e^{-\beta \Delta H(\tau^{\pm})}}{1 + e^{-\beta \Delta H(\tau^{\pm})}} \ , \tag{B1}$$

where $\Delta H(\tau^{\pm})$ measures the change in cost that would be achieved upon accepting the rewiring move. In the above definition, $\beta \geq 0$ plays the role of an "intensity of choice" parameter: for $\beta \to 0$ we have $p \to 1/2$, i.e. the probability becomes independent of the reciprocity target and rewiring moves are accepted or refused at random, while for $\beta \to \infty$ we have

$$p\left(\beta, \tau^{\pm}\right) \to \begin{cases} 1 & \text{if} \quad \Delta H(\tau^{\pm}) < 0 \\ 1/2 & \text{if} \quad \Delta H(\tau^{\pm}) = 0 \\ 0 & \text{if} \quad \Delta H(\tau^{\pm}) > 0 \ , \end{cases} \tag{B2}$$

i.e. rewiring moves are accepted (rejected) as soon as they decrease (increase) the cost function, and moves that do not affect reciprocity are accepted at random.

The link rewiring works as follows.

- Two pairs of distinct nodes (i,k) and (j,ℓ) connected by two links of the same sign (i.e. $A_{ik} = A_{j\ell} = \pm 1$), and such that $A_{i\ell} = A_{jk} = 0$, are chosen at random.
- The change $\Delta H(\tau^{\pm})$ in cost function that would be attained by disconnecting the existing links $i \to k$ and $j \to \ell$ and replacing them with links $i \to \ell$ and $j \to k$ (i.e. setting $A_{ik} = A_{j\ell} = 0$ and $A_{i\ell} = A_{jk} = \pm 1$) is computed.
- With probability $p(\beta, \tau^{\pm})$ given in Eq. (B1) the above rewiring move is accepted and performed. With probability $1 - p(\beta, \tau^{\pm})$ the rewiring move is rejected and all links are kept as they are.
- The above operations are repeated until a steady state is reached, which is ensured by the probabilistic rule (B1), where β plays the role of an inverse temperature in a physical system.

It should be noted that the above rewiring procedure preserves the number of positive/negative ratings received and given by each node i, i.e. it preserves the sums $\phi_i^+ + \gamma_i^+$ and $\phi_i^- + \gamma_i^-$, although, crucially, the individual values of $\phi_i^{\pm}$ and $\gamma_i^{\pm}$ are in general changed. Thus, according to Eq. (A3) the reputation R_i of each node is kept intact, and

All possible rewiring moves are shown in Fig. 1, where we also show when they entail an overall increase in reciprocity ($\Delta\rho^{\pm} > 0$), a decrease in reciprocity ($\Delta\rho^{\pm} < 0$), or when they leave reciprocity untouched ($\Delta\rho^{\pm} = 0$).

Fig. 5 shows the contributions to reputation (introduced in Eq. (1)) obtained by averaging over large samples of null models as functions of both the intensity of choice parameter β and the reciprocity target τ^+. As it can be seen,

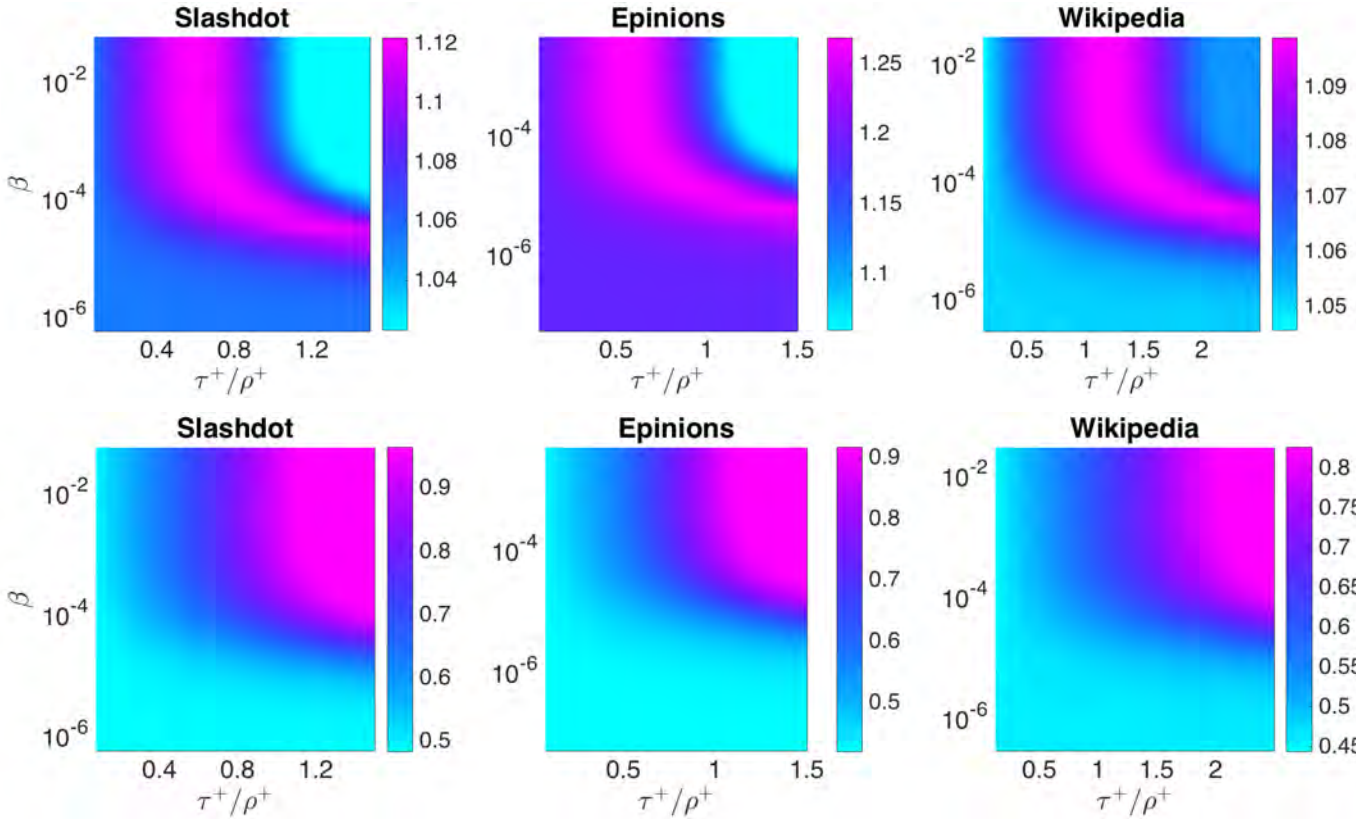

FIG. 5: **Positive reciprocity bias**. The upper panels show the ratio between the average contribution to reputation from unreciprocated positive ratings λ_{Φ}^{+} measured under a null assumption of random link rewiring constrained to preserve each node's reputation and its value in the empirical networks. Lower panels show the ratio between the average contribution to reputation from reciprocated positive ratings λ_{Γ}^{+} measured under the above null assumption and its value in the empirical networks. In each plot such ratio is shown as a function of the intensity of choice parameter β and the reciprocity target τ^+ (normalized by the positive reciprocity ρ^+ measured in the empirical networks). As can be seen, in the empirical networks the contribution to reputation from unreciprocated (reciprocated) positive links is systematically under-expressed (over-expressed) with respect to the null hypothesis.

we systematically find the contribution to reputation from unreciprocated positive links λ_{Φ}^{+} to be under-expressed with respect to any null hypothesis, and, symmetrically, we find the contribution from reciprocated positive links λ_{Γ}^{+} to be over-expressed. As in Fig. 2, λ_{Φ}^{+} has a non monotonic behavior as a function of the reciprocity target, and

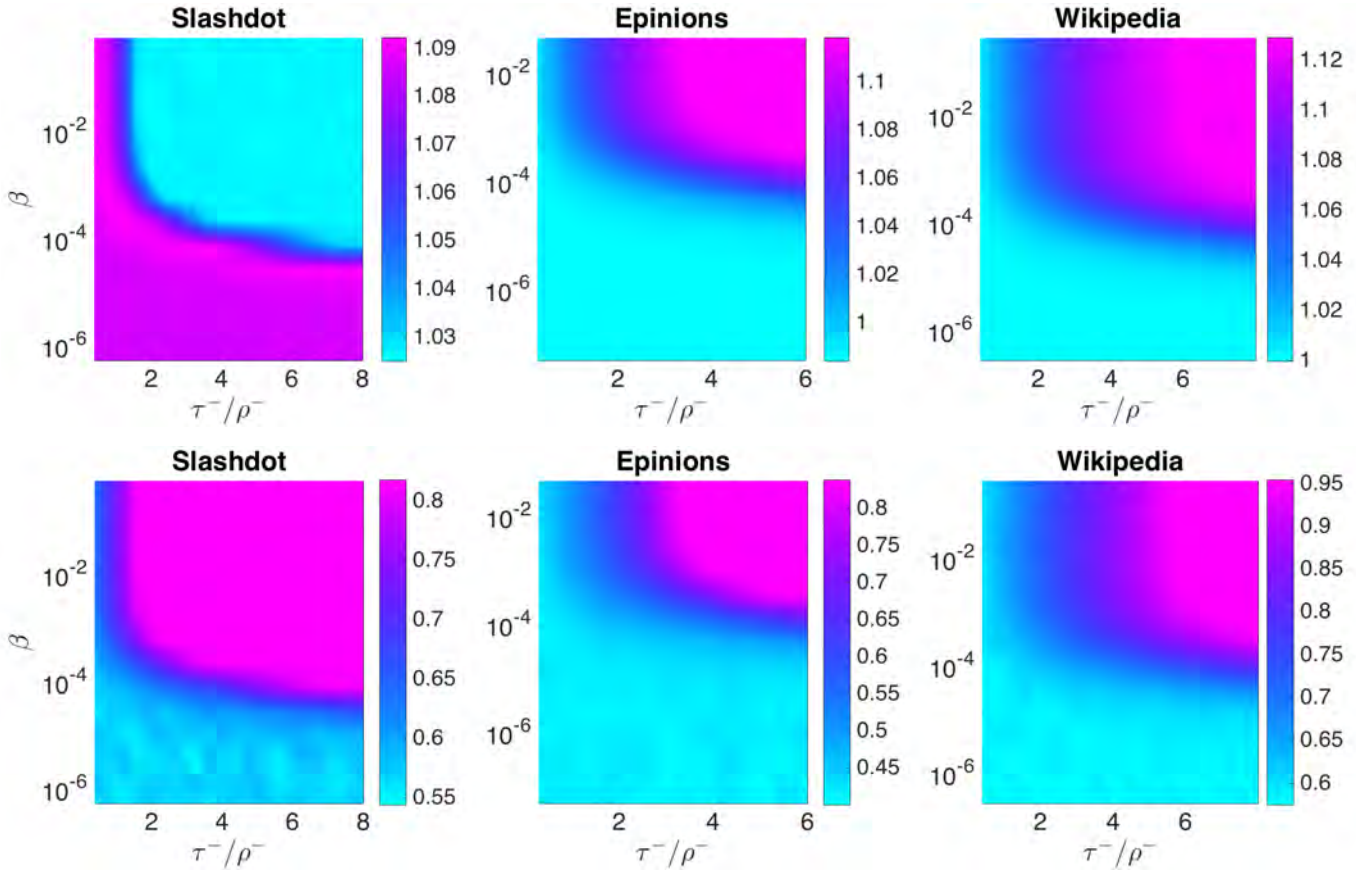

FIG. 6: **Negative reciprocity bias**. The upper panels show the ratio between the average contribution to reputation from unreciprocated negative ratings λ_Φ^- measured under a null assumption of random link rewiring constrained to preserve each node's reputation and its value in the empirical networks. Lower panels show the ratio between the average contribution to reputation from reciprocated negative ratings λ_Γ^- measured under the above null assumption and its value in the empirical networks. In each plot such ratio is shown as a function of the intensity of choice parameter β and the reciprocity target τ^- (normalized by the negative reciprocity ρ^- measured in the empirical networks). As can be seen, in the empirical networks the contribution to reputation from unreciprocated (reciprocated) negative links is systematically under-expressed (over-expressed) with respect to the null hypothesis.

reaches a maximum whose value depends on β, whereas λ_Γ^+ always displays a monotonically non-decreasing behavior both as a function of β and τ^+.

Fig. 6 shows the dependence of the average contributions to reputation from negative ratings. In analogy with the previous case, the contribution from unreciprocated links is systematically under-expressed while that from reciprocated links is systematically over-expressed. In this case, however, Slashdot displays a behavior which markedly differs from the one observed in Epinions and Wikipedia. In fact, in the latter networks both λ_Φ^- and λ_Γ^- are monotonically increasing functions τ^-, which shows that increasing the negative reciprocity target increases both retaliation and constructive negative feedback. Conversely, λ_Φ^- in Slashdot decreases as a function of the target τ^-, i.e. the negative contribution to reputation from constructive feedback decreases as reciprocity is increased. This is suggestive of a possibly different signature of genuinely hostile negative interactions, and suggests that in polarized environments reciprocity should be systematically discouraged.

[1] Lehdonvirta V, Bright J (2015) Crowdsourcing for Public Policy and Government. *Policy & Internet* 7(3):263-267

[2] Van Dijck J, Poell T (2015) Social Media and the Transformation of Public Space. *Social Media + Society* 1(2):2056305115622482

[3] Zervas G, Proserpio D, Byers J (2015) The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry. *Boston University School of Management Research Paper*

[4] Vavilis S, Petković M, Zannone N (2014) A reference model for reputation systems. *Decision Support Systems* 61:147-154

[5] Muchnik L, Aral S, Taylor SJ (2013) Social influence bias: A randomized experiment. *Science* 341(6146):647-651

[6] Salganik MJ, Dodds PS, Watts DJ (2006) Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311(5762):854-6

[7] Zervas G, Proserpio D, Byers J (2015) A first look at online reputation on Airbnb, where every stay is above average. http://papers.ssrn.com/sol3/abstractid=2554500

[8] Lorenz J, Rauhut H, Schweitzer F, Helbing D (2011) How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences of the USA* 108(22):9020-9025

[9] Fehr E, Gächter S (2000) Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives* 19(4):159-181

[10] Bolton G, Greiner B, Ockenfels A (2013) Engineering trust: reciprocity in the production of reputation information. *Management Science* 59(2):265-285

[11] Dowd M (2015) Driving Uber Mad. www.nytimes.com/2015/05/24/opinion/sunday/maureen-dowd-driving-uber-mad.html

[12] Jian L, MacKie-Mason, JK, Resnick P (2010) I scratched yours: The prevalence of reciprocation in feedback provision on eBay. *The BE Journal of Economic Analysis and Policy* 10(1), Article 92

[13] Resnick P, Zeckhauser R, Swanson, J, Lockwood K (2006) The value of reputation on eBay: A controlled experiment. *Experimental Economics* 9(2):79-101

[14] Hu N, Zhang J, Pavlou PA (2009) Overcoming the J-shaped distribution of product reviews. *Communications of the ACM* 52(10):144-147.

[15] Ciotti V, Bianconi G, Capocci A, Colaiori F, Panzarasa P (2015) Degree correlations in signed social networks. *Physica A* 422:25-39

[16] Leskovec J, Huttenlocher D, Kleinberg J (2010) Signed networks in social media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, New York)

[17] Leskovec J, Huttenlocher D, Kleinberg J (2010) Predicting positive and negative links in online social networks. *Proceedings of the 19th international conference on World Wide Web* (Association for Computing Machinery, New York)

[18] Facchetti G, Iacono G, Altafini C (2011) Computing global structural balance in large-scale signed social networks. *Proceedings of the National Academy of Sciences of the USA* 108(52):20953-20958

[19] Heider F (1946) Attitudes and cognitive organization. *Journal of Psychology* 21(1):107-112

[20] Szell M, Lambiotte R, Thurner S (2010) Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences of the USA* 107(31):13636-13641

[21] Altafini C (2013) Consensus problems on networks with antagonistic interactions. *Automatic Control, IEEE Transactions on* 58(4):935-946

[22] Garlaschelli D, Loffredo M I (2004) Patterns of link reciprocity in directed networks. *Physical Review Letters* 93(26):268701

[23] Newman MEJ (2010) Networks: an introduction (Oxford University Press: Oxford)

[24] Taleb NN (2012) Antifragile: Things that Gain from Disorder (Random House: New York and Penguin: London)

[25] Biondo AE, Pluchino A, Rapisarda A (2013) The beneficial role of random strategies in social and financial systems. *Journal of Statistical Physics* 151(3-4):607-622