SCIENTIFIC REPORTS

OPEN

Excess reciprocity distorts reputation in online social networks

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

Received: 29 November 2016
Accepted: 28 April 2017
Published online: 14 June 2017

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 digital reputation, revealing that P2P networks display exceedingly high levels of reciprocity. In fact, these are much higher than those compatible with a null assumption that preserves the empirically observed level of agreement between all pairs of nodes, and rather close to the highest levels structurally compatible with the networks' reputation landscape. This indicates that the crowdsourcing process underpinning digital reputation can be significantly distorted by the attempt of users to mutually boost reputation, or to retaliate, through the exchange of ratings. We uncover that the least active users are predominantly responsible for such reciprocity-induced bias, and that this fact can be exploited to obtain more reliable reputation estimates. Our findings are robust across different P2P platforms, including both cases where ratings are used to vote on the content produced by users and to vote on user profiles.

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. In recent years we have indeed witnessed how 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[4,5]. Given the expected growth of the P2P paradigm, digital reputation will increasingly become central in our online lives, as it will determine access to substantial economic opportunities. Hence, it is crucial to ensure that digital peer-review systems produce reliable reputation scores.

Being decentralised, P2P systems are often thought to promote more economic freedom and democratisation. Yet, their current lack of regulation exposes them to a number of biases which can distort their functioning[6–9]. Game theoretic considerations[10–12], and plenty of anecdotal evidence, suggest that users are often incentivised to reciprocate both positive and negative ratings.

In this paper we show that P2P systems are indeed statistically characterized by excessive reciprocity, and that, on average, one reciprocated rating contributes to a user's reputation more than an unreciprocated one. The fact that reciprocity strongly affects reputation is rather relevant. Indeed, although it is true that the formation of social ties tends to be driven by homophily[13], which in turn might lead similar individuals to reciprocate more than they would do otherwise, it is also well documented that P2P platform users also exchange positive ratings to mutually boost reputation. The "5 for 5" practice of Uber drivers and passengers, i.e. agreeing on exchanging 5 star ratings at the beginning of a ride, is a common firsthand experience of such a practice[14], and similar phenomena have been reported in the interactions between eBay buyers and sellers[15,16] and Airbnb hosts and guests[17]. Symmetrically, negative reciprocity due to mutually honest negative feedback needs to be distinguished from the retaliatory exchange of negative ratings. Therefore, our results show that a central issue in the design of P2P rating systems is that of discerning the information content of reciprocated ratings.

Reputation in online marketplaces is known to affect a buyer's willingness to pay[16]. In this respect, the exchange of ratings effectively introduces *externalities* in P2P platforms[18], as they prevent users from making informed *ex ante* decisions about their peers. Moreover, the anticipation of retaliatory behaviour can discourage users from providing negative feedback[19]. In fact, it is well known that online ratings are often skewed towards positive values[8,20]. In summary, an excess of reciprocity may deteriorate the overall information content in P2P

[1]University College London, Department of Computer Science, 66-72 Gower Street, WC1E 6EA, London, United Kingdom. [2]Systemic Risk Centre, London School of Economics and Political Sciences, Houghton Street, WC2A 2AE, London, United Kingdom. Correspondence and requests for materials should be addressed to G.L. (email: g.livan@ucl.ac.uk)

	N_{tot}	N	L^+	L^-	ξ^+	ξ^-
Slashdot	79,120	4,611	105,115	29,190	4.9×10^{-3}	1.4×10^{-3}
Epinions	131,828	8,732	425,377	40,996	5.6×10^{-3}	5.4×10^{-3}
Wikipedia	138,592	5,538	160,606	29,327	5.2×10^{-3}	9.6×10^{-4}

Table 1. Network statistics. The first column shows the total number of nodes N_{tot} in the original unfiltered networks. The other columns show the number of users N, the number of positive (L^+) and negative (L^-) ratings, and the sparsity $\xi^{\pm}$ (measured as the ratio between the number of existing links and the overall number of possible links, i.e. $\xi^{\pm} = \sum_{i,j=1}^{N} \Theta(\pm A_{ij})/(N(N-1))$ of the networks when restricted to a high participation core of actively engaged users with at least ten combined ratings (received and given).

platforms and pose a threat to their fairness and transparency. We quantify the excess of reciprocity in P2P rating systems with respect to a range of null hypotheses, and measure the impact it has in shaping online reputation. We do so by taking a deliberately reductionist approach. Namely, we investigate three case studies of P2P platforms with binary interactions, i.e. platforms whose users can either choose to endorse or reject their peers' activity. These environments represent a stylised template of the feedback systems underpinning P2P platforms, and allow for a parsimonious representation in terms of signed networks: A positive (negative) interaction between two platform users can be represented as a link carrying a positive (negative) sign between two nodes in a graph. Such systems have attracted considerable attention in the network literature[21–23], as they offer a natural laboratory to test theories for systems with antagonistic interactions, such as social balance theory[24–26] and consensus formation[27].

This network representation is particularly useful to quantify the statistical significance of reciprocity and its impact on user reputation. Indeed, the properties of multi-agent interacting systems can often be encoded into well defined *network motifs*. This, in turn, allows to quantify the statistical significance of empirically observed features by verifying whether certain motifs are still observed in null models in which the network topology is partially randomised. This approach has found successful applications in a great variety of fields, contributing to identifying relevant patterns, e.g., in the world trade web[28] and in interbank credit networks[29].

We encode reciprocity as the fraction of mutual dyads in a network[30], and compare the impact it has in shaping user reputation in the three networks we study as opposed to the one observed in two ensembles of null models. We build a first ensemble by partially randomizing the empirical networks with a link reshuffling procedure designed to preserve the reputation of each user at a predefined level of positive or negative reciprocity. We then proceed to investigate the role of homophily, which, as mentioned above, may act as a natural source of reciprocity, especially in platforms where the nature of positive interactions is particularly friendly, e.g., where positive ratings are expressed by "friending" peers. This is particularly challenging with the data we employ, as they do not provide any categorisation of the nodes in terms of relevant features, which prevents any direct measurement of homophily. To this end, we devise a simple metric of preference similarity to quantify the propensity of pairs of nodes to agree in their endorsements or dislikes of other peers. We adopt such a metric as a proxy for homophily, and build a second null network ensemble through a link reshuffling procedure that preserves this quantity for each pair of nodes together with the reputation of each node. In conclusion, our work cannot explicitly distinguish between "malicious" reciprocity (as in the aforementioned 5 for 5 practice) and "benign" reciprocity, but provides statistically robust results by assessing the likelihood of generating empirically observed features of P2P platforms through partially randomised dynamics.

Within this framework, we quantify the role of rating reciprocity in distorting user reputations, and we eventually demonstrate that more reliable estimates of user reputation can be obtained by discounting reciprocated ratings from the least active users. As we will discuss in detail, the three case studies we analyse share a number of regularities despite the very different nature of the user interactions taking place in each of them.

Results

Signed networks and reciprocity. We analyse data from three P2P platforms: Slashdot, Epinions, and Wikipedia (see *Materials and Methods*). The data are freely available and can be downloaded from the Koblenz Network Collection repository[31]. We apply a filtering procedure to all three networks in order to discard the contributions from casual platform users and only retain the activity of actively engaged ones. In Table 1 we provide details about the size and composition of such networks, and in the *Supplementary Information* we provide detailed evidence that our results do not depend on the specificities of the filtering procedure.

We represent a P2P platform of N users exchanging positive and negative ratings as a signed network, i.e. a set of N nodes described by a square $N \times N$ adjacency matrix A, whose entries are $A_{ij} = 1$ ($A_{ij} = -1$) if user i has given a positive (negative) rating to user j, and $A_{ij} = 0$ if node i has not rated node j. We say that a pair of positive (negative) links are reciprocated if $A_{ij} = A_{ji} = +1$ ($A_{ij} = A_{ji} = -1$). With this notation, one can introduce the number of unreciprocated positive and negative ratings received by a user i (in the following $\Theta(x)$ denotes the step function such that $\Theta(x) = 1$ for $x > 0$ and $\Theta(x) = 0$ otherwise)

$$\phi_i^+ = \sum_{j=1}^{N} \Theta(A_{ji})[1 - \Theta(A_{ij})], \quad \phi_i^- = \sum_{j=1}^{N} \Theta(-A_{ji})[1 - \Theta(-A_{ij})], \tag{1}$$

and the number of reciprocated positive and negative ratings received by a user i

$$\gamma_i^+ = \sum_{j=1}^{N} \Theta(A_{ji})\Theta(A_{ij}), \quad \gamma_i^- = \sum_{j=1}^{N} \Theta(-A_{ji})\Theta(-A_{ij}). \tag{2}$$

In the *Supplementary Information* we provide details about the statistical properties of the above quantities. We then define $\Phi^+ = \sum_{i=1}^{N} \phi_i^+$ ($\Phi^- = \sum_{i=1}^{N} \phi_i^-$) and $\Gamma^+ = \sum_{i=1}^{N} \gamma_i^+$ ($\Gamma^- = \sum_{i=1}^{N} \gamma_i^-$) as the total number of positive (negative) ratings exchanged in the platform that belong to each category.

Within this context, we define positive (negative) reciprocity as the fraction of ratings $i \to j$ that have a matching rating $j \to i$ of the same sign, and we denote it as ρ^+ (ρ^-). With the above definitions we have

$$\rho^+ = \frac{\Gamma^+}{L^+}, \quad \rho^- = \frac{\Gamma^-}{L^-}, \tag{3}$$

where $L^+ = \Phi^+ + \Gamma^+$ ($L^- = \Phi^- + \Gamma^-$) is the total number of positive (negative) links. From a network perspective, this definition is meaningful in the systems in the systems under consideration as they are very sparse (see Table 1). In dense networks, where links are somewhat forced to reciprocate simply due to structural constraints, it could be replaced by the one introduced in[32], which discounts density-related effects. This definition reads $\hat{\rho}^+ = (\rho^+ - \xi^+)/(1 - \xi^+)$, where ρ^+ is as in (3) and $\xi^+ = \sum_{i,j=1}^{N} \Theta(A_{ij})/(N(N-1))$ represents the density of the positive subnetwork. This definition straightforwardly generalizes to the negative subnetwork, and we report the corresponding values in Table 1. As it can be seen from the Table, all density values are of order 10^{-3} or 10^{-4}, and therefore have a negligible impact on reciprocity.

Reputation. Using the quantities introduced in Eqs (1) and (2), we define the reputation of user i as

$$R_i = \frac{L_i^+ - L_i^-}{L_i^+ + L_i^-} = \frac{\phi_i^+ - \phi_i^- + \gamma_i^+ - \gamma_i^-}{\phi_i^+ + \phi_i^- + \gamma_i^+ + \gamma_i^-}, \tag{4}$$

i.e. as the difference between the number of positive ($L_i^+ = \phi_i^+ + \gamma_i^+$) and negative ($L_i^- = \phi_i^- + \gamma_i^-$) ratings received normalised by the overall number of ratings received. The above definition of reputation is such that $-1 \le R_i \le 1$, where $R_i = 1$ ($R_i = -1$) for a user that has received positive (negative) ratings only.

We then define the average contributions to reputation associated with one unreciprocated or reciprocated positive (negative) rating, which we denote as λ_Φ^+ and λ_Γ^+ (λ_Φ^- and λ_Γ^-), respectively. Recalling that $\Phi^\pm$ and $\Gamma^\pm$ indicate, respectively, the total numbers of unreciprocated and reciprocated positive/negative ratings in the networks, and introducing the total reputation in the network $R = \sum_{i=1}^{N} R_i$, we can write

$$R = \Phi^+ \lambda_\Phi^+ - \Phi^- \lambda_\Phi^- + \Gamma^+ \lambda_\Gamma^+ - \Gamma^- \lambda_\Gamma^-.$$

From the above equation one can obtain the following explicit expressions:

$$\begin{aligned} \lambda_\Phi^+ &= \frac{1}{\Phi^+} \sum_{i=1}^{N} \frac{\phi_i^+}{\phi_i^+ + \phi_i^- + \gamma_i^+ + \gamma_i^-} \quad \lambda_\Phi^- = \frac{1}{\Phi^-} \sum_{i=1}^{N} \frac{\phi_i^-}{\phi_i^+ + \phi_i^- + \gamma_i^+ + \gamma_i^-} \\ \lambda_\Gamma^+ &= \frac{1}{\Gamma^+} \sum_{i=1}^{N} \frac{\gamma_i^+}{\phi_i^+ + \phi_i^- + \gamma_i^+ + \gamma_i^-} \quad \lambda_\Gamma^- = \frac{1}{\Gamma^-} \sum_{i=1}^{N} \frac{\gamma_i^-}{\phi_i^+ + \phi_i^- + \gamma_i^+ + \gamma_i^-}. \end{aligned} \tag{5}$$

Excess reciprocity. We compare the reciprocity observed in the empirical networks with the one measured under two null assumptions designed to preserve the overall reputation landscape of the network: In both cases we reshuffle links in the network while preserving the numbers of positive/negative ratings received and given by each individual node (see *Materials and Methods*). We also introduce a positive/negative reciprocity target $\tau^\pm$ and we require the reshuffling algorithms to converge towards it by means of a cost function which depends on an "intensity of choice" parameter $\beta \ge 0$. When $\beta = 0$ the reshuffling procedure is fully random and not sensitive to the reciprocity target. On the other hand, for large values of β the algorithm forces the networks, compatibly with the aforementioned constraints, towards configurations that produce the desired level of reciprocity $\tau^\pm$ (see *Materials and Methods*).

We indicate the two null models as NM1 and NM2, respectively, and we specify them as follows.

- NM1 produces randomised configurations of the empirical networks at a predefined positive (negative) reciprocity target τ^+ (τ^-) while preserving the reputation score R_i (see Eq. (4)) of each node i.
- NM2 produces randomised configurations of the empirical networks at a predefined positive (negative) reciprocity target τ^+ (τ^-) while preserving both the reputation score of each node and the *preference similarity* of each pair of nodes.

We define the preference similarity of a pair of nodes (i, j) as

$$S_{ij} = \sum_{\ell=1}^{N} A_{i\ell} A_{j\ell}. \tag{6}$$

		Null model 1		Null model 2	
	ρ^+	ρ_0^+	ρ_{SAT}^+	ρ_0^+	ρ_{SAT}^+
Slashdot	41.3%	[2.28;2.53]%	[46.6,47.0]%	[4.81;4.96]%	[42.4;42.6]%
Epinions	42.4%	[2.33;2.49]%	[48.2;48.4]%	[4.58;4.66]%	[43.1;43.2]%
Wikipedia	17.6%	[3.01;3.27]%	[36.8;37.2]%	[3.49;362]%	[25.4;25.8]%

Table 2. Over-expression of positive reciprocity in P2P platforms. Comparison between the positive reciprocity ρ^+ observed in the three networks we analyse and the 99% confidence level intervals for the corresponding "basal" levels ρ_0^+ and saturation levels ρ_{SAT}^+ obtained under a null hypothesis of random link rewiring constrained to preserve each user's reputation (null model 1), and a null hypothesis further constrained to also preserve the preference similarity of each pair of nodes (null model 2, see Eq. (6)).

		Null model 1		Null model 2	
	ρ^-	ρ_0^-	ρ_{SAT}^-	ρ_0^-	ρ_{SAT}^-
Slashdot	15.9%	[0.35;062]%	[25.6;26.2]%	[9.05;9.55]%	[20.2;20.8]%
Epinions	7.70%	[1.08;1.42]%	[24.9;25.3]%	[7.33;7.90]%	[18.5;19.0]%
Wikipedia	8.50%	[1.84;2.45]%	[48.4;49.1]%	[8.57;9.18]%	[33.7;34.2]%

Table 3. Over-expression of negative reciprocity in P2P platforms. Comparison between the negative reciprocity ρ^- observed in the three networks we analyse and the 99% confidence level intervals for the corresponding "basal" levels ρ_0^- and saturation levels ρ_{SAT}^- obtained under a null hypothesis of random link rewiring constrained to preserve each user's reputation (null model 1), and a null hypothesis further constrained to also preserve the preference similarity of each pair of nodes (null model 2, see Eq. (6)).

The above quantity measures the level of agreement between two nodes. Indeed, whenever nodes i and j both endorse or dislike a third peer ℓ (i.e. when $A_{i\ell} = A_{j\ell} = 1$ or $A_{i\ell} = A_{j\ell} = -1$) the above sum increases by one, whereas when the two nodes disagree (i.e. when $A_{i\ell} \neq 0$, $A_{j\ell} \neq 0$, and $A_{i\ell} = -A_{j\ell}$) it decreases by one. In this respect, we consider preference similarity as a reasonable proxy for the underlying level of homophily between pairs of nodes. Let us also remark that NM1 preserves preference similarity *globally*, i.e. it preserves the sum $S = \sum_{i<j} S_{ij}$. On the other hand, NM2 preserves preference similarity both locally and globally, and it also captures fairly well the empirical statistical properties of the preference similarity between pairs of nodes that share and do not share reciprocated relationships (see *Supplementary Information*).

We first compare the empirically observed reciprocity levels with the ones measured in the above null models when carrying out the link reshuffling at $\beta = 0$, which produces maximally randomised counterparts of the reputation landscapes observed in the empirical networks, and therefore allows to measure a "basal" reciprocity $\rho_0^{\pm}$ that remains in the system due to its density and the constraints on reputation and preference similarity. Table 2 shows that the positive reciprocity measured in the empirical networks is markedly over-expressed with respect to the positive basal reciprocity levels of both null models. Table 3 reports the results we obtained for negative reciprocity. Slashdot is the only network whose empirical negative reciprocity is over-expressed with respect to both null models. Indeed, while negative reciprocity in Epinions and Wikipedia is substantially over-expressed with respect to NM1, it is compatible with the one measured in NM2 in Epinions, and it is slightly under-expressed with respect to NM2 in Wikipedia.

For large β, we can instead push the reshuffled networks towards targets higher than the reciprocity observed in the empirical networks. 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 still preserving each node's reputation and, in the case of NM2, local preference similarity. We report such values as ρ_{SAT}^+ and ρ_{SAT}^- in Tables 2 and 3, respectively. Both Slashdot and Epinions reach saturation shortly after the target τ^+ exceeds the actual positive reciprocity ρ^+, i.e. the ratio ρ_{SAT}^+/ρ^+ is only slightly larger than one. This is especially evident in NM2 where the local structure of preference similarity is kept intact. On the contrary, Wikipedia can sustain values of reciprocity much larger than ρ^+, i.e. the ratio ρ_{SAT}^+/ρ^+ is larger than 2 in NM1 and close to 1.5 in NM2. We relate this to the different nature of the interactions. Indeed, interactions in Slashdot, where positive and negative links correspond to users tagging each other as "friend" or "foe", encourage backscratching and retaliatory behaviour, whereas a collaborative environment such as Wikipedia is subject to a different incentive structure. This picture is corroborated by the findings on negative reciprocity, where the ratio ρ_{SAT}^-/ρ^- increases substantially as progressing from Slashdot to Wikipedia. The general remark one can make from such results is that more polarised P2P environments are closer to their reciprocity saturation levels.

Production of reputation through reciprocity. We now ask whether reciprocity biases reputation in P2P systems, and, if so, to what extent. To this end, we have divided ratings into four categories: unreciprocated positive ratings, unreciprocated negative ratings, reciprocated positive ratings, and reciprocated negative ratings (see Eqs (1) and (2)). Unreciprocated ratings can be reasonably assumed to represent objective assessments, and

	λ_{Φ}^{+}	λ_{Γ}^{+}	λ_{Φ}^{-}	λ_{Γ}^{-}
Slashdot	0.0323	0.0355	0.0340	0.0522
Epinions	0.0156	0.0220	0.0236	0.0157
Wikipedia	0.0270	0.0325	0.0362	0.0321

Table 4. Evidence that reciprocated ratings contribute more to reputation than unreciprocated ones. $\lambda_{\Phi}^{\pm}$ denote the average contribution to reputation from a positive/negative unreciprocated rating, while $\lambda_{\Gamma}^{\pm}$ denote the average contribution from a positive/negative reciprocated rating.

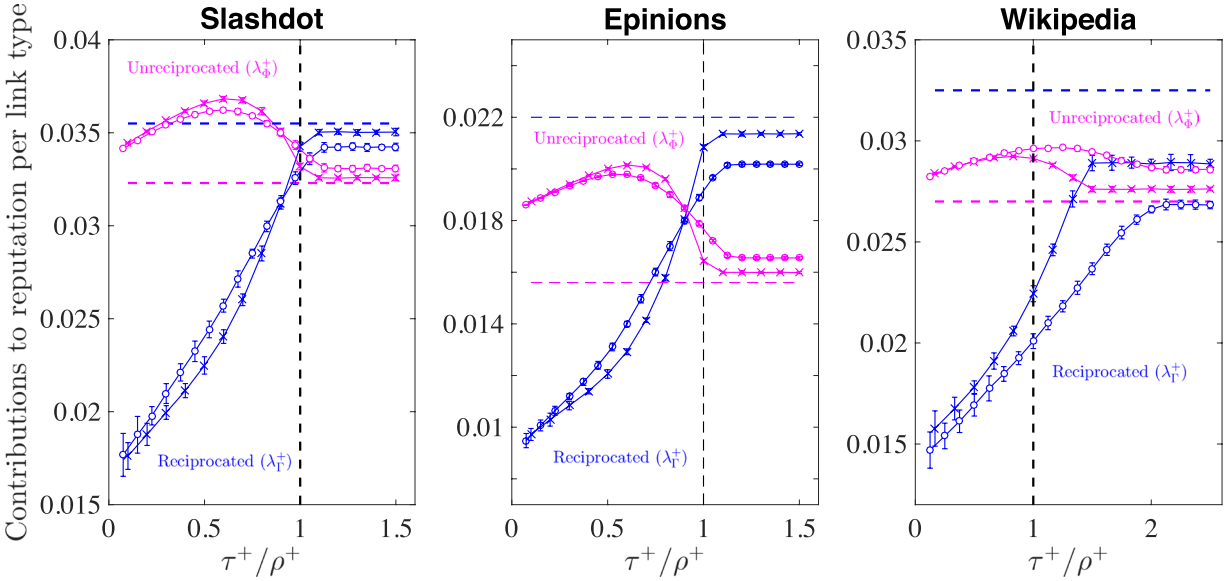

Figure 1. Demonstration that reputation is affected by the reciprocity bias. Behaviour of the average contribution to reputation from unreciprocated positive ratings (λ_{Φ}^{+}, pink) and reciprocated positive ratings (λ_{Γ}^{+}, blue) under two null hypotheses of random link rewiring designed to produce a predefined positive reciprocity target τ^{+}. Circles refer to a null hypothesis constrained to preserve the reputation of each user (null model 1), while crosses refer to a null hypothesis further constrained to also preserve the preference similarity of each pair of nodes (null model 2, see also Eq. (6)). The behaviour 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 1). 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 3). The fact that the contribution from reciprocated (unreciprocated) activity in the actual platforms is systematically lower (higher) than under our null hypotheses highlights the existence of the reciprocity bias.

their contribution to reputation can be thought of as a proxy of a user's "true" reputation. On the other hand, a fraction of the reciprocated ratings could be due to collusion and retaliation.

We compute the average contribution to reputation coming from ratings belonging to each of the four above categories. We do so by means of the quantities introduced in Eq. (5). Table 4 reports the values of these quantities. In all cases we find that $\lambda_{\Gamma}^{+} > \lambda_{\Phi}^{+}$, i.e. on average a reciprocated positive rating contributes more to total reputation than an unreciprocated one. We instead find mixed signatures in the case of negative ratings: 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 ratings play a larger role in damaging reputation than unreciprocated ones. We interpret this as a signature of retaliatory behaviour.

We test the statistical significance of the above findings by resorting again to our two null hypotheses based on constrained random link rewiring. Fig. 1 shows the average contributions to reputation coming from positive reciprocated and unreciprocated ratings as functions of the ratio between the reciprocity target τ^{+} and the positive reciprocity ρ^{+} of the empirical networks. A number of relevant results can be deduced from this Figure.

First, the behaviour of the two quantities as functions of τ^{+} is markedly different. The contribution from reciprocated positive ratings λ_{Γ}^{+} is monotonically increasing, whereas the contribution from unreciprocated positive ratings λ_{Φ}^{+} attains a maximum in correspondence of a certain reciprocity target. In Slashdot and Epinions, such a value is around 50% of the reciprocity observed in the empirical networks in both null models. Conversely, reciprocity in the Wikipedia network has to be increased with respect to its original level in order to reach the maximum contribution from unreciprocated activity.

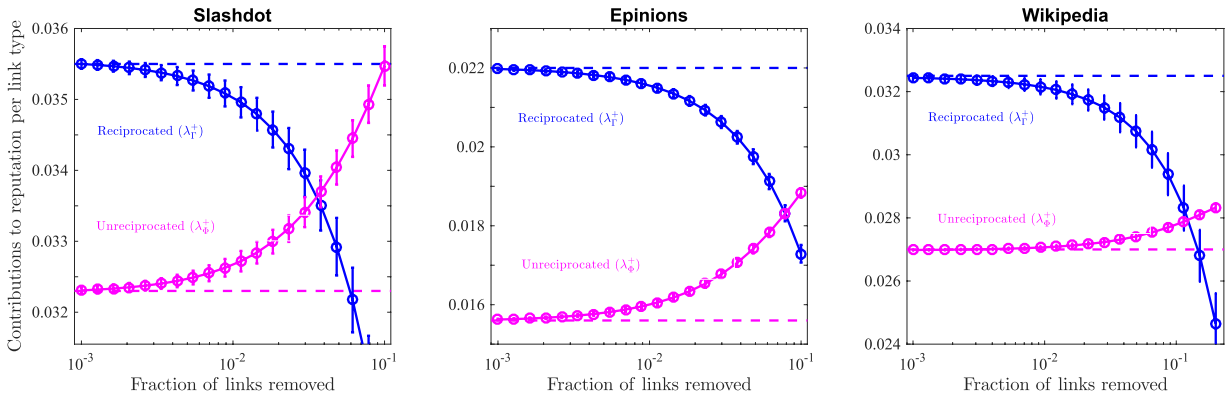

Figure 2. Demonstration that the elimination of a small fraction of ratings suppresses the reciprocity bias. Solid lines show the average contribution to reputation from unreciprocated (λ_Φ^+, pink) and reciprocated (λ_Γ^+, blue) positive links as a function of the fraction of reciprocated positive links removed from the network. Circles represent the behaviour 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. Error bars represent 99% confidence level intervals.

Second, throughout the whole range of reciprocity shown in Fig. 1, we find that the contribution to reputation from unreciprocated activity is under-expressed in the real systems. Symmetrically, the contribution from reciprocated ratings is systematically over-expressed. Both these conditions hold regardless of the specific null model (see Section S5 in the *Supplementary Information* for a comparison with a different null model based on link and sign reshuffling), and hold for any value of the parameter β (see Fig. S5 in the *Supplementary Information*). This highlights the existence of what we call *reciprocity bias*: Reciprocated activity plays an exceedingly large role in shaping reputation at the aggregate level.

Third, in null models the relative importance between reciprocated and unreciprocated activity is reversed with respect to the one observed in the actual networks. As shown in Table 4, in the empirical networks one positive reciprocated rating always contributes more to reputation, on average, than an unreciprocated one (i.e. $\lambda_\Gamma^+ > \lambda_\Phi^+$). In contrast, under our null hypotheses 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. Notably, both in Slashdot (for NM1) and Wikipedia one has $\lambda_\Gamma^+ < \lambda_\Phi^+$ even when the reciprocity target is kept equal to its empirical value, i.e. when $\tau^+ = \rho^+$ (dashed vertical lines in Fig. 1). This is a case where our null hypotheses entail 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. In the following we further demonstrate that P2P dynamics drive the networks towards very "atypical" states whose main features are not robust to small perturbations, and we will exploit this point to investigate possible prototypes of regulatory countermeasures to prevent users from building reputation through excessive reciprocity.

Suppressing the reciprocity bias through random link elimination. Figure 2 shows the behaviour of the average contribution to reputation from reciprocated (λ_Γ^+) and unreciprocated (λ_Φ^+) positive ratings upon the removal of reciprocated ratings. Namely, we randomly select pairs of users and check whether a reciprocated rating between them exists. If so, we remove it. Notably, in Slashdot the deletion of 3% of the reciprocated positive ratings (i.e. slightly more than 1.2% of the overall positive ratings) is enough to make the contributions to reputation from reciprocated and unreciprocated ratings statistically compatible. The same result is achieved by removing roughly 8% of the reciprocated positive ratings in Epinions, and roughly 11% of the reciprocated positive ratings in Wikipedia (corresponding, respectively, to 3.3% and 1.9% of the overall positive ratings). Furthermore, one can also see from Fig. 2 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 ratings is lost extremely fast, i.e. by removing roughly 1% of the reciprocated positive ratings.

Given the heavy tailed nature of the distributions of ratings given and received by each node (see Fig. S2 in the *Supplementary Information*), a statistical feature common to many networked systems[33], the above protocol mostly targets and penalises users with a lower number of ratings, while leaving users with many reciprocated links relatively untouched. Although this might seem unfair at first, let us remark that the networks we analyse are high participation cores, where the contribution from casual platform users has already been filtered out. Moreover, such a protocol is meaningful from the viewpoint of user incentives: Indeed, a newcomer to a platform is more incentivised to reciprocate in order to boost her visibility in the network, whereas a high activity user with good reputation has little marginal gain from an additional rating. In this respect, removing ratings between low activity users is key to suppressing the reciprocity bias and improving the average quality of ratings (see Fig. S8 in the *Supplementary Information*).

The above exercise obviously neglects the complexities that translating such a procedure into an actual platform management policy would entail. Indeed, an explicit policy of random link removal of reciprocated links would discourage users from rating each other at all: no rating would be reciprocated (as any reciprocated pair

would become liable to being removed), which in turn would discourage providing ratings in the first place. Yet, it is interesting to consider the above procedure as a thought experiment to measure the fragility of the reciprocity bias.

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 Slashdot, Epinions, and Wikipedia. We have done so by investigating a range of null models designed to preserve the individual reputation of each user and the preference similarities of pairs of users 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 null benchmarks. Moreover, we find that reciprocated ratings contribute more to reputation than unreciprocated ones. This is at odds with what we observe in the aforementioned null models. Even when we incorporate the tendency of users with similar like/dislike patterns to also like each other, we still 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 networks is 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, 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[34–38,40], and suggests that an effective policy to prevent users from building reputation through excessive reciprocity would be that of injecting a small amount of randomness into the system. We validated this hypothesis by carrying out a random link elimination procedure in the three networks we analysed, which shows that the removal of reciprocated links between users with a low number of ratings, hence highly incentivised 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. 1) 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.

Due to the lack of information about the users' identity and features, our paper does not investigate directly how homophily (i.e. the tendency to interact with similar individuals in social networks[13,39]) might contribute to explain the exceedingly high levels of reciprocity and their impact on user reputation. Yet, the null network ensembles employed to carry out our statistical analyses are specifically designed to preserve the preference similarity structure of the networks, i.e. the tendency of certain pairs of users to endorse/reject the same content, which in itself represents a fairly close proxy for network homophily. In particular, we tested whether our conclusions are robust when compared to a null hypothesis that only preserves preference similarity at a global level as opposed to a more stringent null hypothesis designed to preserve preference similarity at the level of individual pairs of nodes. Interestingly, we found our results on positive excess reciprocity and its impact on reputation to be largely independent of the particular null assumption, i.e. the basal positive reciprocity levels and the contributions to reputation (see Fig. 1) observed when keeping the local preference similarity structure intact are not significantly different from those observed when this is not preserved. This suggests that homophily alone would not be able to justify the high levels of positive reciprocity we empirically observe and their impact on user reputation. On the other hand, preserving the local similarity structure essentially captures most of the empirical positive reciprocity. Indeed, although not statistically compatible, the saturation reciprocity levels measured when accounting for preference similarity are remarkably close to the empirically observed ones in Slashdot and Epinions. This suggests that homophily drives the growth of opinion-driven platforms in such a way as to maximize positive reciprocity. Symmetrically, the basal levels of negative reciprocity measured when accounting for homophily essentially capture the empirical levels in Epinions and Wikipedia, i.e. the exchange of negative ratings cannot be distinguished from a partially randomized dynamics when negative interactions lack an explicitly hostile trait.

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 "stylised-yet-complex" nature that we chose signed networks as templates for P2P systems. In this respect, our work advances the existing literature on reciprocity and reputation in online environments, where most empirical

results are tied to the specificities of a particular platform or rating system, and therefore lack generality. Our work provides a "one-fits-all" network methodology that can be used as soon as user interactions in a platform can be classified as either positive or negative, which can be achieved quite easily for the most common types of online feedback (e.g., by thresholding in the case of graded scales and via sentiment analysis in the case of textual ratings). The potential of such a reductionist approach is highlighted by the consistency of the regularities that we detect in the three platforms we analyse, despite the vastly different natures of the user interactions that characterise them. For instance, we observe qualitatively similar patterns of reputation formation in Slashdot, where votes are expressed directly on a whole user profile, and in Wikipedia, where edits express indirect votes on the quality of the content produced by others.

Our analyses show that P2P rating systems are plagued by biases. 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.

Materials and Methods

Data. The data we analyze belong the the following platforms:

- **Slashdot**, a website whose users post, read and comment news about science and technology. The data we analyse pertain to the Slashdot Zoo, i.e. the set of friendship/enmity relationships that Slashdot users form. Namely, users can label each other as "friend" ("foe") in order to endorse (oppose) opinions and activity on the platform.
- **Epinions**, a platform for crowdsourced consumer reviews whose users exchange insight about a variety of products. Based on their review history, users can express trust/distrust relationships to each other.
- A collection of actions from 563 **Wikipedia** articles about politics, where each interaction between users (such as co-edits, antagonistic edits, reverts, restores, etc.) was interpreted as positive or negative value depending on its nature.

A substantial portion of the users in the above networks are casual users who do not interact frequently with the platform. In fact, 32% of Slashdot users, 47% of Epinions users, and 49% of Wikipedia users have either given or received just one rating. We therefore proceed to filter the noisy contribution from casual platform users, in order to ensure that reputation scores are computed from the ratings of actively committed users. The above network datasets do not provide information about possible repeated interactions between the users, which prevents from applying network statistical validation techniques that explictly account for the heterogeneity in user activity (see, e.g. ref. 40). Hence, we choose to restrict the networks to a high participation core of actively engaged users with a number of total ratings (received and given) equal or higher than a threshold t. Clearly, after the deletion of these nodes, some other nodes might result to be disconnected in the restricted network, as there is no guarantee that the t (or more) received/given ratings are exchanged with other nodes within the core. We therefore carry out a second filtering in order to remove disconnected nodes. This operation leaves us with the networks whose statistical properties are described in Table 1. In Section S7 of the *Supplementary Information* we provide evidence that our main results are consistent across different thresholds.

Null models. In order to assess the statistical significance of the features observed in the empirical networks, we define two ensembles of null network models (labeled as NM1 and NM2 in the following) that depend on two parameters. Namely, we define a positive reciprocity target τ^+, and we introduce a cost function $H(\tau^+) = [L^+(\rho^+ - \tau^+)]^2$ to measure the distance between the current positive reciprocity in the network and the target (the following can be straightforwardly generalized to the case of negative reciprocity). 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

$$p(\beta, \tau^+) = \frac{e^{-\beta \Delta H(\tau^+)}}{1 + e^{-\beta \Delta H(\tau^+)}}, \tag{7}$$

where $\Delta H(\tau^+)$ measures the change in cost that would be achieved upon accepting the rewiring move (i.e. the difference between the cost function after and before the rewiring), and $\beta \geq 0$ is an "intensity of choice" parameter that determines the rewiring procedure's responsiveness to changes in cost: When $\beta = 0$ the above probability is equal to 1/2, which amounts to a fully random rewiring (independently of the target τ^+), whereas when β is large rewiring moves that cause an increase (decrease) in cost get rejected (accepted) with probability close to one.

The link rewiring works as follows. In NM1:

- Two pairs of distinct nodes (i, k) and (j, ℓ) connected by two positive links (i.e. $A_{ik} = A_{j\ell} = 1$), and such that $A_{i\ell} = A_{jk} = 0$, are chosen at random.
- The change $\Delta H(\tau^+)$ 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 $A_{i\ell} = A_{jk} = 1$) is computed.

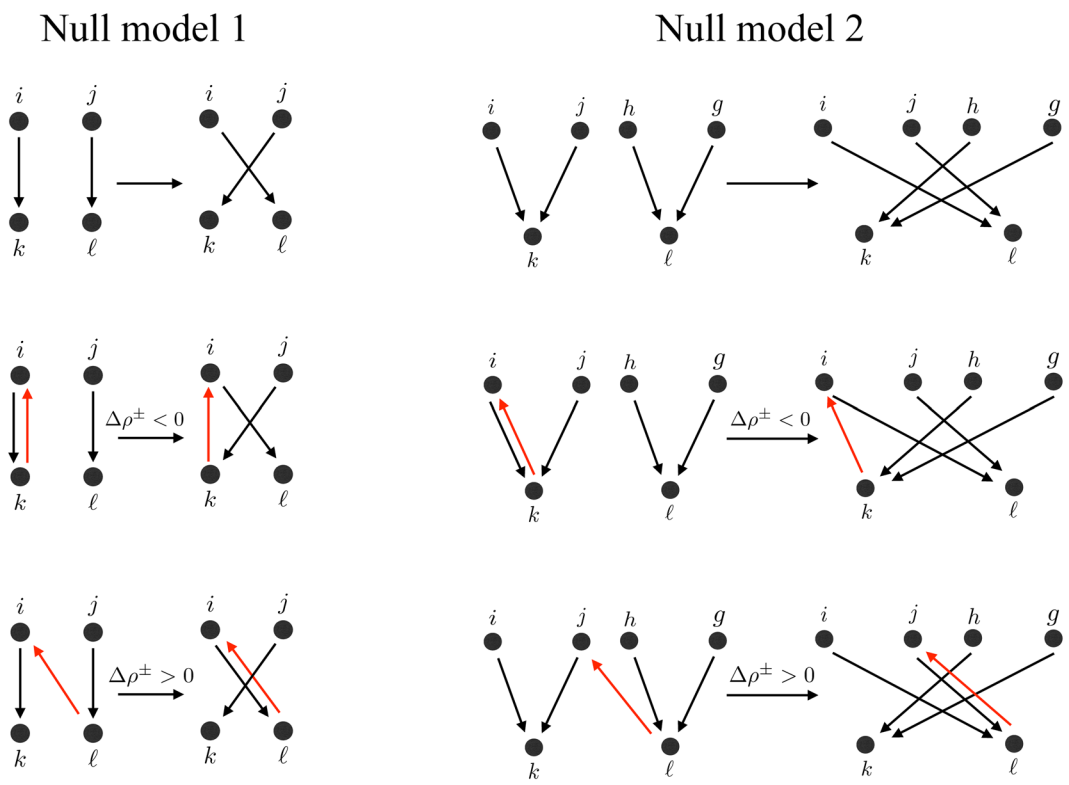

Figure 3. Rewiring moves of the null models. On the top of the Figure we show the fundamental rewiring moves of the two null models we consider. In the central and bottom part of the Figure we also show examples where the rewiring moves contribute to decrease and increase the overall reciprocity, respectively. We highlight the links responsible for such increase/decrease in red.

In NM2:

- Two triplet of nodes (i, j, k) and (h, g, ℓ) connected by at least two pairs of positive links (i.e. such that $A_{ik} = A_{jk} = 1$ and $A_{h\ell} = A_{g\ell} = 1$), and such that $A_{i\ell} = A_{j\ell} = A_{hk} = A_{gk} = 0$, are chosen at random.
- The change $\Delta H(\tau^{+})$ in cost function that would be attained by disconnecting the existing links $i \to k$, $j \to k$ and replacing them with links $i \to \ell$ and $j \to \ell$, and by disconnecting the existing links $h \to \ell$, $g \to \ell$ and replacing them with links $h \to k$ and $g \to k$ (i.e. setting $A_{ik} = A_{jk} = A_{h\ell} = A_{g\ell} = 0$ and $A_{i\ell} = A_{j\ell} = A_{hk} = A_{gk} = 1$) is computed.

Then, for both null models:

- With the probability $p(\beta, \tau^{+})$ in Eq. (7) the above rewiring moves are accepted and carried out. With probability $1 - p(\beta, \tau^{+})$ the rewiring moves are rejected and all links are kept as they are.
- The above operations are repeated until a steady state is reached, which in the long run is ensured by the probabilistic rule in Eq. (7), where β plays the role of an inverse temperature in a physical system (see Section S2 in the *Supplementary Information*).

In Fig. 3 we sketch the two fundamental rewiring moves of the above null models, and we report two examples in which they cause the overall positive/negative reciprocity of the network to decrease ($\Delta\rho^{\pm} < 0$) and to increase ($\Delta\rho^{\pm} > 0$).

The above procedures are reminiscent of the directed configuration model from the literature on complex networks[41,42], and consist in randomly redirecting ratings, hence destroying correlations between raters and ratees, while preserving both the reputation of each node and the system's heterogeneity at a fixed level of reciprocity. In fact, the above rewiring procedures preserve the number of positive/negative ratings received and given by each node i, i.e. they preserve 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. (4) our rewiring procedures keep the reputation R_i of each node intact.

References

1. Lehdonvirta, V. & Bright, J. Crowdsourcing for Public Policy and Government. *Policy & Internet* **7**(3), 263–267 (2015).
2. Van Dijck, J. & Poell, T. Social Media and the Transformation of Public Space. *Social Media + Society* **1**(2), 2056305115622482 (2015).
3. Zervas, G., Proserpio, D. & Byers, J. The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry http://papers.ssrn.com/sol3/papers.cfm?abstract-id=2366898 (Date of access: 29/03/2017) (2016).

4. Tadelis, S. The Economics of Reputation and Feedback Systems in E-Commerce Marketplaces. *IEEE Internet Comput.* **20**, 12–19 (2016).
5. Vavilis, S., Petković, M. & Zannone, N. A reference model for reputation systems. *Decis. Support Systems* **61**, 147–154 (2014).
6. Muchnik, L., Aral, S. & Taylor, S. J. Social influence bias: A randomized experiment. *Science* **341**(6146), 647–651 (2013).
7. Salganik, M. J., Dodds, P. S. & Watts, D. J. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**(5762), 854–6 (2006).
8. Zervas, G., Proserpio, D. & Byers, J. A first look at online reputation on Airbnb, where every stay is above average http://cs-people.bu.edu/dproserp/papers/airbnbreputation.pdf (Date of access: 29/03/2017) (2015).
9. Lorenz, J., Rauhut, H., Schweitzer, F. & Helbing, D. How social influence can undermine the wisdom of crowd effect. *Proc. Natl. Acad. Sci. USA.* **108**, 9020–9025 (2011).
10. Fehr, E. & Gächter, S. Fairness and retaliation: The economics of reciprocity. *J. Econ. Perspect.* **19**, 159–181 (2000).
11. Cimini, G. & Sánchez, A. How evolutionary dynamics affects network reciprocity in Prisoner's Dilemma. *J. Artif. Soc. Soc. Simulat.* **18**(2), 22 (2015).
12. Bolton, G., Greiner, B. & Ockenfels, A. Engineering trust: reciprocity in the production of reputation information. *Manage. Sci.* **59**, 265–285 (2013).
13. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.* **27**, 415–444 (2001).
14. Dowd, M. D U Mad. *The New York Times* http://www.nytimes.com/2015/05/24/opinion/sunday/maureen-dowd-driving-uber-mad.html (2015) (Date of access: 29/03/2017).
15. Jian, L., MacKie-Mason, J. K. & Resnick, P. I scratched yours: The prevalence of reciprocation in feedback provision on eBay. *B. E. J. Econ. Anal. Policy* **10**, Article 92 (2010).
16. Resnick, P., Zeckhauser, R., Swanson, J. & Lockwood, K. The value of reputation on eBay: A controlled experiment. *Exp. Econ.* **9**, 79–101 (2006).
17. Fradkin, A., Grewal, E., Holtz, D. & Pearson, M. Bias and reciprocity in online reviews: Evidence from field experiments on Airbnb. *Proc. 16th ACM Conference on Economics and Computation* (2015).
18. Nosko, C. & Tadelis, S. The limits of reputation in platform markets: An empirical analysis and field experiment http://faculty.haas.berkeley.edu/stadelis/EPP.pdf (Date of access: 29/03/2017) (2015).
19. Dellarocas, C. & Wood, C. A. The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Man. Sci.* **54**, 460–476 (2008).
20. Hu, N., Zhang, J. & Pavlou, P. A. Overcoming the J-shaped distribution of product reviews. *Commun. ACM* **52**, 144–147 (2009).
21. Ciotti, V., Bianconi, G., Capocci, A., Colaiori, F. & Panzarasa, P. Degree correlations in signed social networks. *Physica A* **422**, 25–39 (2015).
22. Leskovec, J., Huttenlocher, D. & Kleinberg, J. Signed networks in social media. *Proc. SIGCHI Conf. Human Factors in Computing Systems* 1361–1370 (2010).
23. Leskovec, J., Huttenlocher, D. & Kleinberg, J. Predicting positive and negative links in online social networks. *Proc. 19th Intl. Conf. World Wide Web* (2010).
24. Heider, F. Attitudes and cognitive organization. *J. Psychol.* **21**, 107–112 (1946).
25. Szell, M., Lambiotte, R. & Thurner, S. Multirelational organization of large-scale social networks in an online world. *Proc. Natl. Acad. Sci. USA.* **107**, 13636–13641 (2010).
26. Facchetti, G., Iacono, G. & Altafini, C. Computing global structural balance in large-scale signed social networks. *Proc. Natl. Acad. Sci. USA.* **108**, 20953–20958 (2011).
27. Altafini, C. Consensus problems on networks with antagonistic interactions. *IEEE T. Automat. Contr.* **58**, 935–946 (2013).
28. Fagiolo, G., Squartini, T. & Garlaschelli, D. Null models of economic networks: the case of the world trade web. *J. Econ. Interact. Coord.* **8**(1), 75–107 (2013).
29. Squartini, T., van Lelyveld, I. & Garlaschelli, D. Early-warning signals of topological collapse in interbank networks. *Sci. Rep.* **3**, 3357 (2013).
30. Wasserman, S. & Faust, K. *Social network analysis: Methods and applications* (Cambridge University Press, 1994).
31. Kunegis, J. KONECT - The Koblenz Network Collection. *Proc. Int. Web Observatory Workshop* 1343–1350 (2013).
32. Garlaschelli, D. & Loffredo, M. I. Patterns of link reciprocity in directed networks. *Phys. Rev. Lett.* **93**, 268701 (2004).
33. Caldarelli, G. *Scale-free networks: Complex webs in nature and technology* (Oxford University Press, 2007).
34. Mantegna, R. N. & Spagnolo, B. Noise enhanced stability in an unstable system. *Phys. Rev. Lett.* **76**, 563–566 (1996).
35. Pluchino, A., Rapisarda, A. & Garofalo, C. The Peter principle revisited: A computational study. *Physica A* **389**, 467–472 (2010).
36. Biondo, A. E., Pluchino, A. & Rapisarda, A. The beneficial role of random strategies in social and financial systems. *J. Stat. Phys.* **151**, 607–622 (2013).
37. Taleb, N. N. Antifragile: Things that Gain from Disorder (Random House, 2012).
38. Albert, R., Jeong, H. & Barabási, A. L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
39. Currarini, S., Jackson, M. O. & Pin, P. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica* **77**(4), 1003–1045 (2009).
40. Li, M.-X. *et al*. Statistically validated mobile communication networks: the evolution of motifs in European and Chinese data. *New J. Phys.* **16**, 083038 (2014).
41. Newman, M. E. J. *Networks: An Introduction* (Oxford University Press, 2010).
42. Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. Critical phenomena in complex networks. *Rev. Mod. Phys.* **80**, 1275–1335 (2008).

Acknowledgements

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

Author Contributions

All authors conceived the research and designed the analyses. G.L. conducted the analyses. All authors discussed the results and wrote the paper.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-03481-7

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.


Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2017