Rise and Fall of reputation in a Web of Trust: the Bitcoin-OTC market case

Gargiulo Floriana[1], Bertazzi Ilaria[2] and Huet Sylvie[2]

[1] *CNRS, GEMASS, 59-61 rue Pouchet, 57017 Paris, France*

[2] *IRSTEA, 9 Avenue Blaise Pascal, 63170 Aubière, France*

ABSTRACT

In this paper we study the dynamics of reputation of people with a particular attention to the cascade phenomena giving rise to rapid falls of credibility. To do this we study the Bitcoin-OTC website: a peer to peer marketplace for trading bitcoin crypto-currency. This platform has a unique characteristics, making it the most adapted playground for analyzing reputation dynamics: a web of trust is implemented containing direct ratings among couples of users. The web of trust is a network where nodes are the users and the weighted links are the evaluations among the users. We analyze the structure and the dynamics of this network with a multilayer approach distinguishing the rewarding and the punitive behaviors. The aggregate values of reward and punishment and the resulting reputation are unequally distributed among the users, generating few users with a very high or low reputation and several with moderate values. The interaction between the layers is not trivial due to the presence of several users with both high rewards and punishments. We characterize the reputation trajectories identifying prototypical behaviors associated to three classes of users: trustworthy, untrusted and controversial. Controversial users are the only ones presenting up and down reputation trends. We focus on them for understanding which are the possible factors driving reputation falls and rise, and which dynamical patterns characterize these cascades: some users have real oscillating behaviors, other abuse of the trust system doing a few good transactions to gain reputation for cheating the users afterwards, other naturally and slowly die out after a long series of positive exchanges (like disappearing from the system) and finally, some users are hardly beaten by organized trolling attacks.

Keywords: reputation, trust, multilayer networks

ISSN 2332-4031; DOI 10.34962/jws-75

1 Introduction

Reputation is a key issue in human society, intimately connected to social inequalities. In real life, reputation is connected to social status and it could be potentially inferred by the socio-demographic indicators describing an individual. In this sense, offline reputation is quite static, being connected to long lasting human characteristics (class, education, gender,...). In web 2.0 reputation assumes an even more central role and new characteristics: it is measurable and more volatile. Web 2.0, and in particular sharing economy that is it's economic structure, are characterized by free peer to peer interaction among "strangers"; for this reason trust models are needed to grant mutual trust between buyers and sellers that have no direct interactions: this trust is built on the experience of other users Resnick *et al.*, 2000; Kollock *et al.*, 1999; Thierer *et al.*, 2015. Online reputation is (more explicitly than real life reputation) a social construct, being the result of a large set of peer evaluation actions (recorded by the platforms). As a direct consequence, online reputation is measurable. In online platforms users start with the same initial condition of null reputation, without any status bias, and their actions drive the reactions of the other users and consequently the dynamics of their reputation. Given the high number of potential users' interactions in online platforms, reputation can change fast.
In online platforms, reputation has a double role: it can be used as a source of information (for future trades) and as a source of potential punishment. In such a way reputation systems are ad-hoc structures to protect platforms from cheating behaviors: trusted behaviors are encouraged to avoid retaliation (this is the concept of the "shadow of the future" Axelrod, 1984) as it has been observed in several papers in game theory Cuesta *et al.*, 2015 and agent based models Conte and Paolucci, 2002; Manzo and Baldassarri, 2015.

All the most important online platforms implement different forms of reputation systems. In most of the cases user-objects interactions are considered Scellato *et al.*, 2011, like in Booking or Trip Advisor O'Connor, 2008, where users rate services. A similar architecture is present in Q&A websites, like StakOverflow, where users evaluate the answers, and not the user that gave the answer. In these contexts the reputation is calculated as the appreciation of the whole user's activity Bosu *et al.*, 2013; Movshovitz-Attias *et al.*, 2013.

In other cases, like Wikipedia, where the activity of the editors is more difficult to be observed and evaluated, a large debate is still present on the possible way to implement reputation measures Adler and De Alfaro, 2007; Zeng *et al.*, 2006.
Few cases are really based on peer to peer interactions where users directly evaluate each other. These systems are the most interesting in order to observe the construction of reputation as an emerging interaction process. In particular several studies on Epinions Richardson *et al.*, 2003; Guha *et al.*, 2004; Leskovec *et al.*, 2010 and the Coachsurfing Lauterbach *et al.*, 2009 platforms showed the central role of reciprocation in these evaluation systems.

Remarkable reputation growth patterns can be explained by the Merton's concept of cumulative advantage Merton, 1988. On the contrary reputation fall cascades, that are observed in online social platforms, cannot be conceptualized in a known theoretical framework. For this reason we focus our study on the reputation falls and, in particular, we aim to observe and characterize these patterns and to understand the triggering factors. This study demands extremely rich data containing time-dependent information on the relative ratings of a set of users. Most of the available datasets on peer to peer rating systems (Epinions, Advogato, Slashdot) cannot be used to analyze the fine dynamics of users' reputation: data do not have a fine temporal resolution or do not contain the full history of the rating system.
For this reason we focus on the only dataset that, to our knowledge, present all the necessary information: the Bitcoin-OTC market platform "https://bitcoin-otc.com" n.d. The details of the database are presented in section 2. Using tools from complex network science, and in particular a new multilayer representation of the web of trust graphs, we show that reputation fall cascades are much faster than reputation growth patterns, confirming experiments showing that reputation is easier to loose than to get Yaniv and Kleinberger, 2000. This is discussed in section 3A. Finally, combining quantitative and qualitative analysis of relevant reputation cascade patterns, we show that different types of endogenous and exogenous factors (cheating behaviors, trolling attacks, etc) can generate different users' "fading" processes. This is discussed in section 3B.

2 Data and methods

Bitcoin-OTC (https://bitcoin-otc.com) is a peer to peer (over-the-counter) online marketplace for trading bitcoin crypto-currency and common goods with bitcoin crypto-currency. To mitigate the risks of the p2p unsupervised exchanges, a Web of Trust is implemented to have access to the counterpart's reputation before a transaction, as presented above. In this web of trust system, a user i can rate another user j with an integer score s_{ij} varying from -10 to 10. This information is publicly available on the the website without any restriction. Starting from the summary page containing the users information (https://bitcoin-otc.com/viewratings.php), we collected all the user-names. Secondly, performing a loop on the users, we crawled all the json files containing the ratings received by all the users, with the associated time-stamps, having a daily resolution. Aggregating these files we reconstruct the whole web of trust of the platform. The dataset contains 5,878 users and 35,795 ratings exchanged between 2011 and 2017. 89% of the ratings has an associated text describing the motivation of the given score. We did not perform an automatic treatment of these texts but we manually analyzed their content in cases of particular controversies.

We assume that the socio-psychological micro-mechanisms governing rewarding and punitive ratings could be different. Due to this reason we study the web of trust as a multiplex weighted directed network with two different layers: the rewarding layer, L^+, containing only the positive scores and the punitive layer, L^-, containing the negative scores.

Each user is therefore identified with a node both on the rewarding and the punitive layer. On both the layers the weighted edges are labelled with the time of the interaction and with the absolute value of the score associated ($w_{ij} = s_{ij}$ for the rewarding layer and $w_{ij} = -s_{ij}$ for the punitive).

As usual in Webs of Trust Guha *et al.*, 2004; Leskovec *et al.*, 2010, the number of edges on the rewarding layer ($N_e^+ = 32305$) is much higher than on the punitive one ($N_e^- = 3490$). As we can observe in Fig.1, also the distribution of the scores is different between the two layers: in L^+, due to the norm suggested by the website, the score $s = 1$ is dominant, while in L^- in order to emphasize the punitive gesture the score $s = -10$ is the most frequent.

To each node i we associate the following values:

- **In-degrees=number of received scores on the two layers**
$$\vec{k_{in}}(i) = (k_{in}^+(i), k_{in}^-(i)) \quad (1)$$
- **Out-degrees=number of given scores (activity) on the two layers**
$$\vec{k_{out}}(i) = (k_{out}^+(i), k_{out}^-(i)) \quad (2)$$
- **Rewards (ρ^+) and Punishments(ρ^-) received**
$$\rho^+(i) = \sum_j w_{ji}^+, \rho^-(i) = \sum_j w_{ji}^- \quad (3)$$

Finally, for each user we define the **global reputation**, as it is reported and visible for all the users on the website:

$$\rho(i) = \rho^+(i) - \rho^-(i) \quad (4)$$

Notice that users have, at a first glance, the information on the aggregate reputation from the user summary page, but they also have a more detailed view of the whole history of the ratings, if they are interested in a more precise evaluation of other users.

These indicators reconstruct a static aggregation of the rating process, namely the photography of the trust system at the final state of the evolution: the punishments and the rewards are defined by the received ratings on each label, summed together independently from the time they were given.
A finer temporal description can be given constructing the daily reputation scenarios:

$$\rho^{+(-)}(i,t) = \sum_j \sum_{t' \le t} w_{ji}^{+(-)}(t') \quad (5)$$

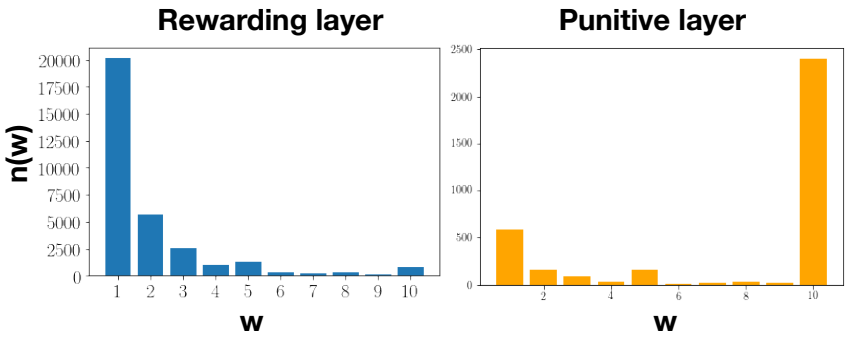


Figure 1: Distribution for the rewarding (left plot) and the punitive (right plot) layer

3 Results

3.1 Hard to climb, easy to fall

3.1.1 The relationship between reward and punishments

In Fig2A we analyze the probability distributions of the rewards and punishments (Eq.3). We notice that, notwithstanding the significant differences between the score distributions and the number of edges of the two layers, the positive and negative reputations follow the same power law distribution ($P(\rho^{+(-)}) \sim (\rho^{+(-)})^{-2}$). In the inset, we displayed the distribution of the global reputation ρ of the users. The global reputation is definitely not symmetric, showing non-trivial interactions between the two layers.

Clustering coefficient measures the tendency to form triangles in a network structure: for a node i, a clustering coefficient $c(i) = 1$ signifies that in its neighborhood all the triples of nodes are connected as a triangle, $c(i) = 0$. On the contrary, implies a star topology (without triangles). The clustering coefficient of the nodes as a function of the total degree (for the undirected and unweighted version of the network) is displayed in Fig.2B. We can observe that, for the rewarding layer, the clustering is in general higher than for the punitive layer, above all for high degree nodes. Comparing each layer with a randomized reshuffling of the network that maintains the degree distribution (configuration model), we observe that the clustering coefficient is higher than the null case for the rewarding layer and lower for the punitive. This is coherent with the balance theory suggesting that, in a triadic structure, three negative interactions, are not balanced (Heider, 1944; Antal *et al.*, 2006).

After the statistical properties of the network we concentrated on the ranking of the nodes according to the different attributes. In particular we concentrated to the values of $k^+_{in}, k^-_{in}, k^+_{out}, k^-_{out}$ and ρ. In Fig.2C we represented these quantities according to the ranking for k^+_{in}, the number of positive ratings received. The nodes with a high in-degree in the rewarding layer usually are the most active (high out-degree both on the rewarding and the punitive layer), and clearly have an high reputation. Notice however that the nodes higher in the k^+_{in} ranking also receive several negative scores (k^-_{in}).

3.1.2 Trustworthy, untrusted and controversial

We analyze now the properties of the users between the two layers, and in particular we analyze the position of the users in the space (ρ^+, ρ^-). We divide the plane in three areas, as described in Fig3A: $A1 = [\rho^- < 0.5\rho^+]$, $A2 = [0.5\rho^+ < \rho^- < 2.0\rho^+]$, $A3 = [\rho^- > 2.0\rho^+]$.

The users in the first area, $A1$, have an high level of rewards level and a low level of punishments, therefore, in this area we can place the *trustworthy* users. On the contrary the users in the third area, $A3$, are *untrusted*, having a low level of rewards and an high level of punishments. Finally, the users in the second area, $A2$, are *controversial* having quite similar values for rewards and punishments. The largest part of the users are trustworthy. In the higher plot of Fig.3B we can observe the distribution of the global reputations in the three different areas. Not surprisingly the untrusted users have a negative reputation and the trustworthy ones a positive one. The reputations associated to the controversial are lower. More interestingly, we can observe that the untrusted users have in general a lower activity ($\vec{k_{out}}$) on both the layers. The controversial users have a similar activity on the two layers and in general have the highest activity on the punitive layer. Finally the trustworthy users are extremely active for rewarding and much less for punishing. We can argue that the untrusted users are like "trolls" appearing and cheating one or more users just one time. In such a way they fast construct their negative reputation and after disappear. Controversial users are real users that gain and give negative scores according to more complex mechanisms that better mimic the formation of reputation in the human society. In this sense, their activity on the punitive layer could be interpreted as a reciprocation of one or more negative scores.

3.1.3 Linear and complex trajectories

We will now analyze the trajectories of individuals' reputations, namely the dynamical evolution of reputation in time (Eq.4). In particular we analyze, for each user, the time flattened trajectories: the sequence of reputation changes (dropping the times when reputation does not change). In the left plot of Fig.4 we show the trajectories for the users that, at a certain point of the system evolution, entered the top 10 of rewards

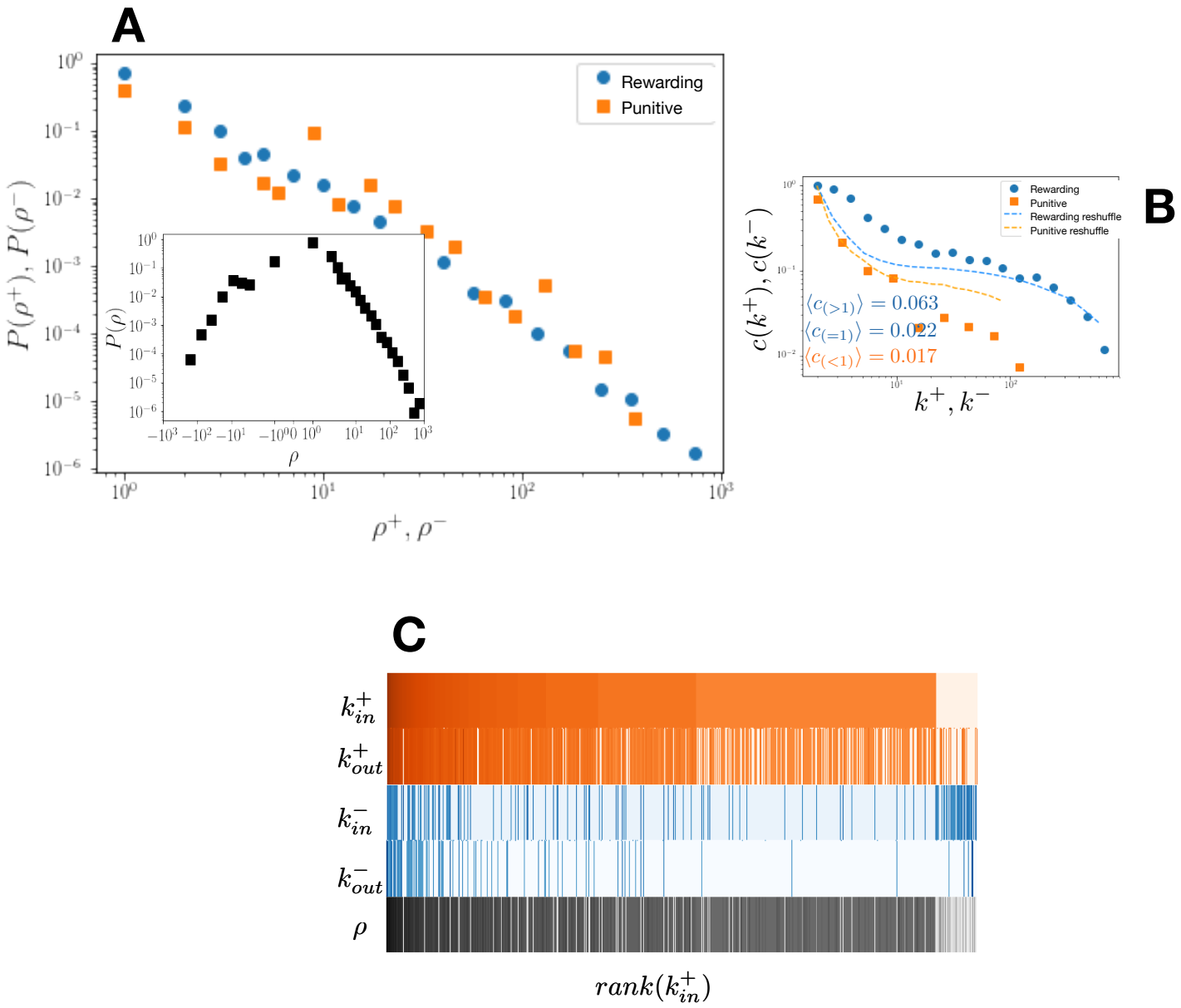


Figure 2: Static properties. Plot A: distribution of the rewards and punishments. A logarithmic binning procedure is applied to flatten the tails of the distributions. Inset: Global distribution of the reputation. Plot B: Clustering coefficient spectrum for the rewarding and the punitive layer, compared with the spectrum for a randomized version of the network. Plot C: In and out degrees of the nodes on the two layers and reputation, with the nodes sorted according to the in-degree ranking: the higher values on the left. The color intensity is proportional to the indicator value.

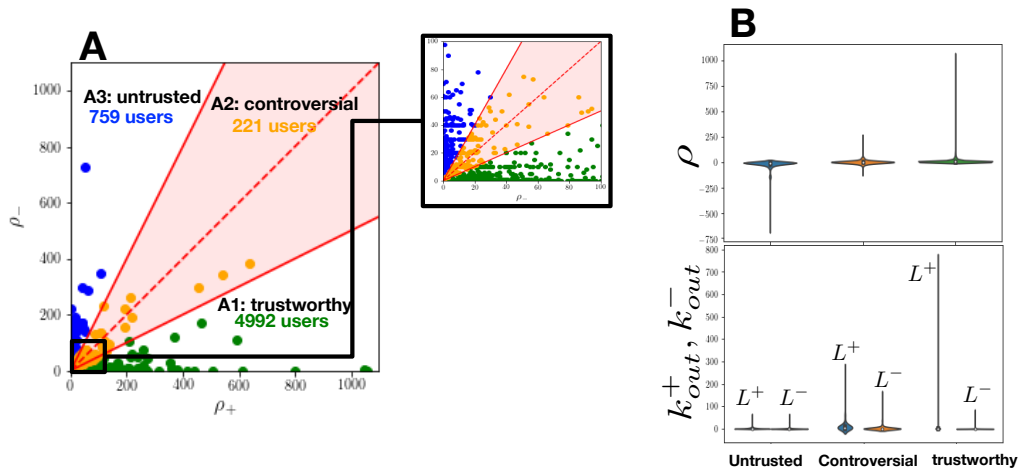


Figure 3: Plot A: Identification of trustworthy (green), untrusted (blue) and controversial users (orange) with a zoom on the lower values. Plot B: Violin plot of the normalized distributions of the reputation (upper plot) and of the activities (lower plot) for the three categories of users.

or punishments. Notice that while the users in the top 10 of rewards have a linear growth, with a slope usually higher than the linear one, users in the top 10 of punishments, can in general have high global reputations since they alternate phases of growing rewards and phases of growing punishments. In the right plot of Fig.4 we show the trajectories for trustworthy, untrusted and controversial users, as defined previously. Untrusted users have fast linear reputation decreasing trajectories while trustworthy users have slow linear increasing trajectories. Controversial users, experience phases of reputation growth following the slow linear trend of trusted users alternated to fast reputation crashes, following the fast linear trend of untrusted users. These trajectories, and in particular the behavior of controversial users, that alternate growing and decreasing phases, confirm a well known fact in economic science, but never observed before in peer to peer evaluation systems: reputation is hard to get and easy to loose.
In the next section we will focus on controversial users, in order to understand which are the social mechanisms that can trigger an inversion of the reputation growth, generating rapid decreasing cascades.

3.1.4 Summary

- Both the reward and the punishment are unequally distributed among the users. Moreover a non-trivial interaction exists between the punitive and the rewarding layer, given by the fact that several highly rewarded users are also often punished.
- We can identify three categories of users: *trustworthy*, that are much more rewarded than punished, *untrusted*, that are much more punished than rewarded, and finally *controversials* that are both punished and rewarded.
- Analyzing the users' trajectories of scores, we showed that getting a good reputation is an hard and slow process while loosing it is a fast and abrupt process.

3.2 The focus on controversial users

While the dynamics of trustworthy and untrusted users can be easily understood in terms of cumulative (dis)advantage, the patterns characterizing the sequences of rise and falls of the controversial users result much more interesting. Which are the determinants of the inversions? Are they due to the users' commercial behaviors, to external factors or to a hidden organization in the social network? For studying and categorizing the controversial patterns we first analyze at the global level the dynamics of the punitive layer and after, using the information extracted by this analysis, we enter into the details of the individual controversial trajectories.

3.2.1 The anatomy of negativity

In this paragraph, we will focus on the punitive behavior analyzing the temporal evolution of the punitive layer. We define the "negativity" of the system, at time t, as the sum of the negative scores for $t' < t$. Similarly, "positivity" is the sum of the positive scores. In Fig.5A we observe that while positivity has a smooth growth, negativity undergoes sudden transitions in few moments; such slopes are consistent with results coming from Bertazzi *et al.*, 2018. In general jumps of bitcoin value increase the trading activity of the website and consequently the activity of the web of trust, due to new encounters. The activity picks on the positive lay are strictly related to bitcoin value. However activity picks on the negative layer, not related to the bitcoin value, were detected. We identify, at first, these transition points as the points where the derivative of the negativity function explodes. We name these transition moments as "attack days". We justify such label in relationship to the comments we will shortly present. The list of attack days is reported in Fig.5B. Notice that some of these points are located in temporal phases where also positivity grows (namely phases of high traffic on the website), however the largest jumps correspond to phases of 'normal' activity where the derivative of negativity grows but the derivative of positivity remains constant. In the following we will focus on these days where only negativity grows.

To better understand the system behavior during the attacks, we analyze the size of the giant component of the punitive layer, day by day (Fig.5C). For the largest part of time the largest component has size $S_g = 2$, meaning that, at most, isolated couples interact. The attack days identified by the derivative of negativity perfectly correspond to situations where the giant component has large size $S_g > 20$ indicating more complex interaction patterns: not several independent couples but combined interactions between a subgroup of users.

In Fig.6 we analyze the topological properties of the day by day networks, separately focusing on the cases $S_g = 2$ (Fig.6A),$S_g = 3$ (Fig.6B),$S_g > 20$ (Fig.6C). In the cases of dyadic interactions we focus on the reciprocity issue, namely we count how the fraction of cases where, after a negative score between i and j, a corresponding negative score is reported between j and i. This is motivated by the qualitative analysis of the comments associated to the negative ratings, often reporting, as the reason for a negative score, the fact of having received a negative score from the other user:

> «when i get it, ill get to it»

We observe that however only 24.7% of the negative scores are followed by a reciprocation feedback. Analyzing the text associated to the scores, we can observe that a significant part of the negative ratings can be associated to social imitation processes:

> «NOBODY LIKES YOU»
> or
> «never spoke to or traded the guy »

In the case of triadic interactions the first important observation is that triangles are never formed. Most activities are conjoint attacks (two-against-one) or bi-attacks (one-against-two). The high percentage of conjoint attacks confirms the fact that a "social" organized activity exists behind this website.
Finally we analyze the structure of the network in the 'attack days', corresponding to the cases where $S_g > 20$. For doing

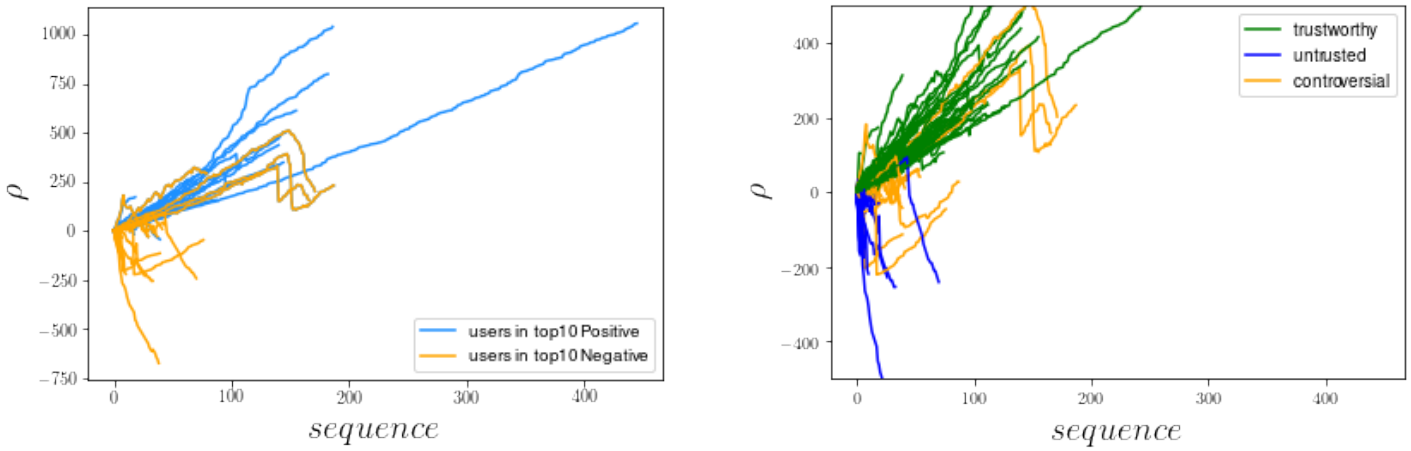


Figure 4: Left plot: Flattened reputation trajectories for users in the top 10 list of rewards and punishments. Right plot: Flattened reputation trajectories for trustworthy, untrusted and controversial users.

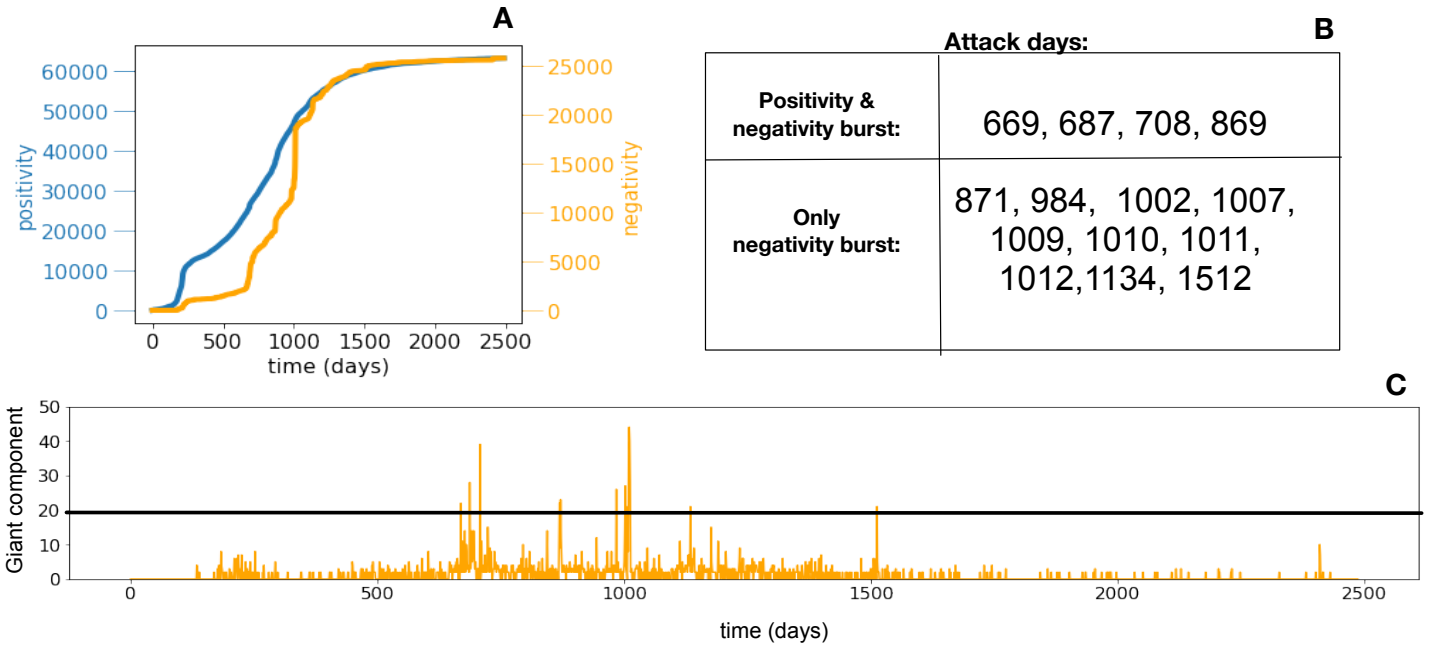
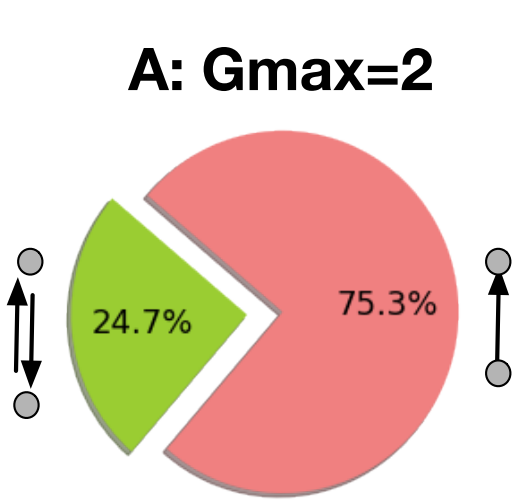
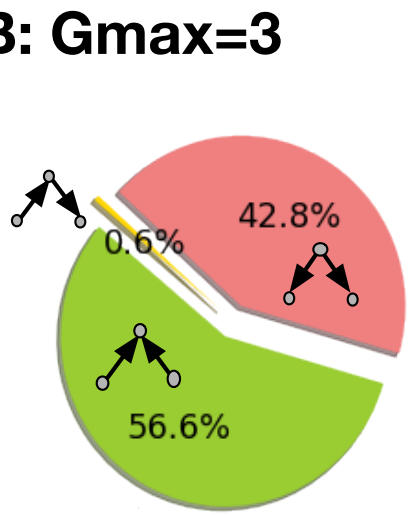


Figure 5: (A, upper left - B, upper right, C - bottom)
Plot A. Positivity and Negativity as a function of time.
Table B. Attack days identified by the analysis of the derivative of negativity function and by the giant component size.
Plot C. Size of the giant component of the negative layer.





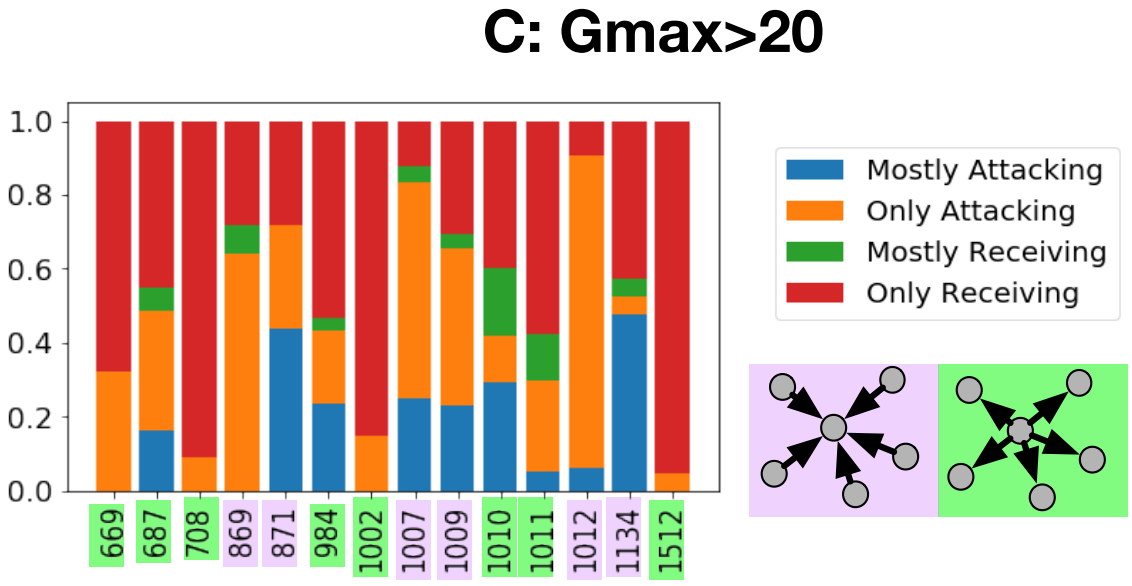


Figure 6: (A, upper left - B,upper right, C- bottom)
Plot A: Fraction of reciprocated links in cases where S_g=2.
Plot B: Percentage of motif structures for S_g=3. Notice that triangles are never present.
Plot C: Ego networks characterization in the attack days. We call the nodes with only outgoing links 'only attacking', with only incoming links 'only receiving', with more outgoing than incoming 'mostly attacking' and with more incoming than outgoing 'mostly receiving'. Days reported in green are those with more cases of single user attacking several other users and vice versa, while those reported in violet correspond to cases where it prevails a single user is attacked by several others.

it we study the shape of the ego networks for each user for each days, categorizing these in 4 categories: pure outgoing stars (one user attacking several others), named 'only attacking' users, pure incoming star (several users attacking one), named 'only receiving', and the two intermediate categories more outgoing links than incoming ('mostly attacking') and viceversa ('mostly receiving'). We represent the abundance of each category for each day. In general the pure categories are more present than the others. The most diffused case corresponds to attacks done by few users to several others. However we also observe, during some days (1007,1009) a real coordination between groups of users somehow engaged in a sort of "clan war".

Analyzing the texts and the nick names (usually the same name followed by a number) of the involved users we can easily understand that these particularly intense days correspond to a misuse of the web of trust where a group of individuals create fake account to attack (and after counter attack) other groups with a typical trolling behavior. This trolling behavior can be explicitly identified by the users names (anonymized in the figure), for example, in the network reported in Fig.7, at time $t = 871$, where the T and the AT users represent groups of accounts with small variations of the same nick name attacking each other. Second rounds of negative ratings are usually accompanied by comments like:

> «anyone with the time to rate 297 people negatively has got to be the scum of the Earth»
> or
> «Sock account of scammer/spammer *userXXX*
> »

which may appear to be a retaliation collective process from misuse of rating system and creating several fake identities. In the network graphs presented in Fig.7, the size of the nodes indicates the indegree (so a large node is heavily attacked) and the color the reputation (red low reputation, blue high reputation). Interesting behaviors can be observed in the sequence $t = 1002, 1007, 1010$. Massively attacked users (i.e. U102) react with several attacks in the following times and attacked back. Note that a clear correlation between the attacking/receiving patterns and reputation at the attack times (color of the nodes) does not explicitly appear. In this sense a sort of oscillation of the network edges' direction (from incoming to outgoing star shapes) appears in the temporal evolution around the most central attack days.

3.2.2 How we can fall

In the previous section we showed that the temporal evolution of the negative layer presents topological jumps that could strongly drive the evolution of the reputation trajectories. We will now use this information to better understand the individual trajectories of the controversial users, focusing on the 49 users that received more than 10 ratings. We first categorize the reputation time series of the controversial users $\rho_i(t)$ using a hierarchical clustering procedure. The clustering is calculated applying, as metric distance, the time series correlation. The obtained dendogram structure, identifying five classes of users, is presented in Fig.8. For each class we plot the average value and the standard deviation for each day. The dotted lines represent the attack days. Users in green and yellow clusters are clearly strongly affected by the attacks around day 1000. Those in the green cluster had previously gained a high reputation that suddenly dropped out with the attacks and that remained stationary after. Users in the yellow cluster had a low reputation becoming negative after the attacks. Users in the red cluster present oscillatory growing and decreasing reputation after the attacks. The cyan cluster represents user whose reputation grew strongly before the first attack ($t = 871$) and smoothly lost a bit until a stable point. Finally, violet users represent quite low reputation users losing their reputation after the strongest attack days.

To conclude some of the controversial behaviors, those that are related to the attack days (yellow and green), represent situations where users suddenly loose reputation due to a coordinate action of the other users.

Yellow users seam to firstly take advantage of some good interactions and then, after the loss of reputation, they try to regain their position. Some commentaries on this type of users before this attack of others look like:

> «Acted as escrow agent for a deal – very professional!»
> «quick transaction»
> «Trusted buyer! Hope to deal with him in the future again! :D»

Then, they look to be punished from a group of users genuinely disappointed by their behavior, like:

> «Impulsive. Doesn't seem serious about paying past debts.»
> or
> «joining the bandwagon»

After being attacked, they look like trying to behave correctly from time to time, possibly to regain credibility. They may receive comments like:

> «traded a little amount of LTC for BTC with him - smooth transaction»

Violet and cyan users are not that different, except that they do not suffer from a sudden attack, but they loose reputation smoothly. They reached a certain level of reputation (quite low for violet larger for cyan) and loose it slowly day by day due to a final gradual withdrawn from the system. They shifted from moderately positive votes commentaries like:

> «Based in how he handled a dispute on 27May2012, I would do business with him.

to negative rates with not so harsh, credible commentaries, like:

> «Late debt payment on BTCJAM https://btcjam.com/listings/5
> Will remove if repaid.»
> «I don't trust these guys. Something fishy happened.»

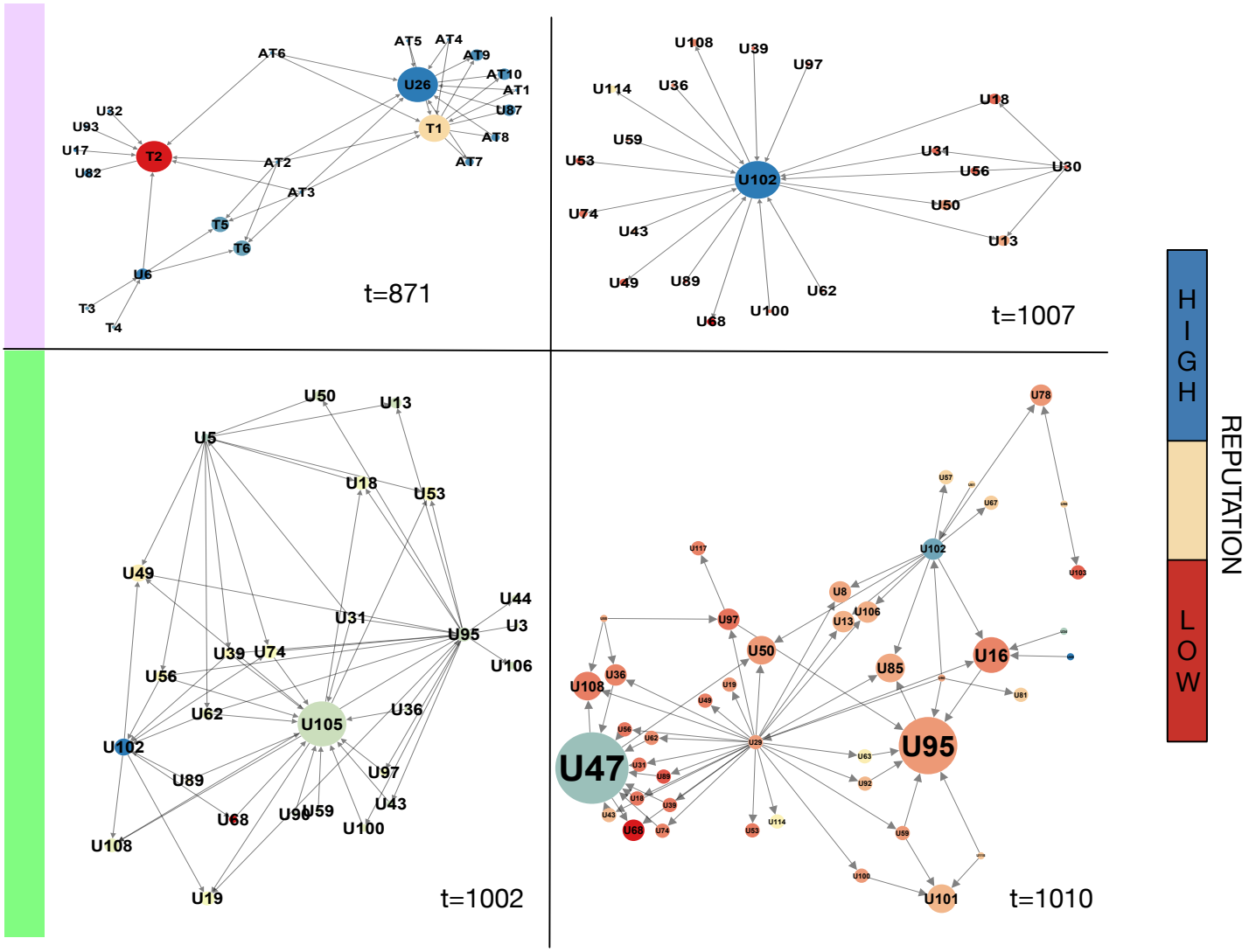


Figure 7: Anonymized network structures in some attack days (t=871,1002,1007,1009). Upper plots correspond to network structures with a mostly incoming star topology, while lower plots with a mostly outgoing star. The color of the nodes represent the reputation (red: low, clue: high). The size of the node is proportional to the in-degree.

The difference between the last two groups lies in the final state: violet type end in a negative total reputation, cyan end anyway somehow positive, meaning that they leave the market before being totally discredited. There is the possibility that they didn't have the chance to do business again and the community simply stopped downgrading them. Users in the violet clusters are usually short trajectories representing users 'exploiting' the market system with few positive transaction to gain reputation followed by a list of negative behaviors.

Red and green users, on the other hand, do not look like receiving credible commentaries when attacked being rated badly. They look like having a smooth growth in reputation at the beginning, then they are particularly attacked (in a specific moment for most of green users, in different time for red ones), with commentaries that take harsher tones, many being evidently copy-paste of others', like (we avoid reporting swears, which where numerous) :

> «SCAMMER!! BEWARE »
> «DONT TRUST »
> «SCAMMMEMRMERMERMEMREMRMERM»

It is not clear weather such attack where due to bad negotiations or discussions which triggered an emotional response from rating users, or unjustified trolling.

Red users are somehow resilient to such attacks, they continue to do business on the market and slowly regain in reputation. Green users, on the other hand, never recover from the loss of credibility and supposedly left the market.

3.2.3 Summary

- The temporal evolution of the negative layer presents topological jumps that strongly influence the evolution of the reputation trajectories.
- These jumps correspond to organized trolling activity, followed up by coordinate reactions
- The temporal trajectories of reputation loss are different in case of of trolling attacks and in case of variation in the perception of the user's behavior (from perceived as acting correctly to perceived as cheating). In the first case the fall is abrupt, in the second one much smoother.

4 Conclusions - Discussion

In this paper, we aim to understand how and why reputation of people rise and fall. The study of the data of the website Bitcoin-OTC shows that three types of evolution of reputation are noticeable. Two of them are "pure" evolutions: one almost always increasing, that we call the trustworthy users' evolution, and another one almost always decreasing, that we call the untrusted users' evolution. The third one alternates increasing and decreasing reputations phases; we call it the evolution of controversial users.

In any case, looking at the slope of the decreasing and increasing phases, we notice that the decrease is stronger than the increase. In other words, our results show that reputation is gained slowly while it is lost sharply. This result has been already proven in a slightly different off-line context with the psychological experiments of Yaniv and Kleinberger, 2000 and Bonaccio and Dalal, 2006. In particular, this studies prove that good reputation is more easily lost than gained, and authors have argued about some complementary explanations for this phenomenon, like the idea of asymmetry of trust Slovic, 1993, or the heuristics used for an impression judgment: judgments are inordinately influenced by an actor's more negative attributes. This is probably related to the fact a negative information is perceived as more diagnostic of an actor's true character than positive information is Skowronski and Carlston, 1989.

Then our question is finally which factors define and moderate the slope of the decrease compared to the slope of the increase of the reputations. Our investigations of global properties of reward and punishment on the one hand, and of the temporal trajectories of evolution of the reputation, especially the reputation of controversial user, point out first results. We describe and discuss them in the following, distinguishing what possibly comes from human behaviours and biases, from what comes from the design of the Bitcoin-OTC website.

4.1 Explanations linked to human behaviours

We showed rewards are more frequent but punishments are more intense. This makes good reputation frequent but very sensitive to punishment.

It has been often argued that people put more positive scores than negative one to avoid retaliation. Reciprocity, and its negative counterpart, retaliation, have been shown existing in human relationships, especially in internet marketing Resnick *et al.*, 2000. However, other explanations can be considered. Firstly this can reflect that the majority of people are honest, respecting the norms, implying satisfaction of buyers is largely more common than dissatisfaction. Secondly, it has been shown by Dellarocas and Wood, 2008 that unsatisfied buyers tend to remain silent: "eBay traders are more likely to post feedback when satisfied than when dissatisfied".

Complementary to investigate why negative feedbacks are rarer, one can wander why they are so intense, a lot more intense than rewards. Several explanations can be proposed.

At first, we have seen in our data that there are several troll attacks using strongly negative scores, and collective reactions of protection against trolls, that explained the use of the strongest negative score by trolls and anti-troll people. Then, it points out that strong emotional reactions due to troll-threat lead to intense negative scores. But some others issues are at stake in the importance of the negative score, especially:

- lack of integrity is much more punished that lack of competence Yin *et al.*, 2010. This is probably the way our trolls are perceived;
- in the maintenance of organizational norm process, Whitson *et al.*, 2015 have shown that "observers consistently punished more than direct recipients did and that direct recipients rewarded more than observers did . . . observers

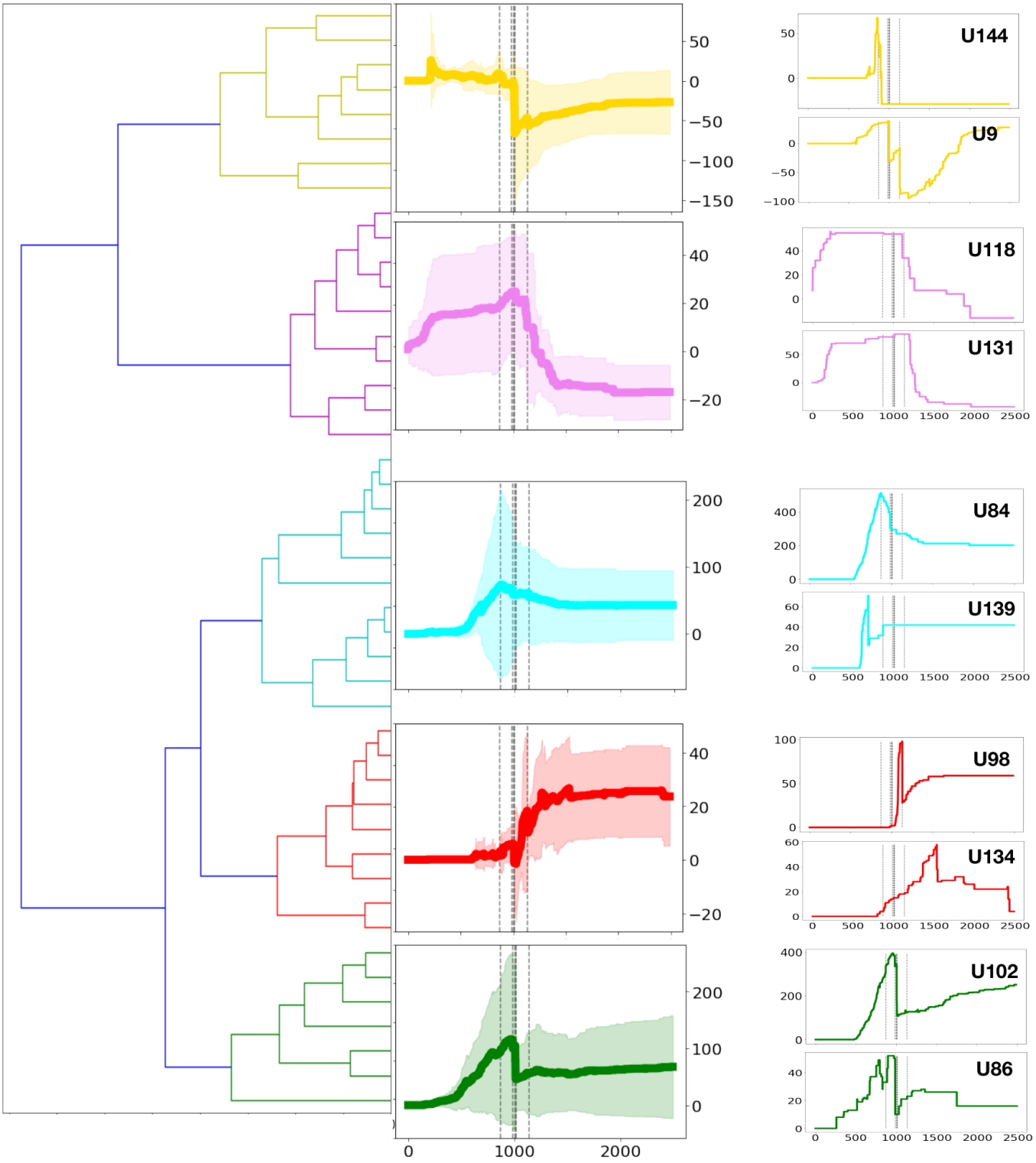


Figure 8: Dendogram of the users trajectories and average behaviors by cluster. The central plots represent the average trajectory of the users in the cluster with the error (shadowed area). The right plots represent some typical trajectories.

felt a stronger obligation to punish but a weaker obligation to reward than recipients did“. This represents the reaction of our anti-trolls agents which, by the way, are able to give strong scores without any transactions, thanks to the website's rules.

We have also observed temporal “controversial users“ trajectories decreasing in reputation less sharply, day-by-day. Specific scores of such decrease have not been studied. However, ordinary falls of reputation have been studied elsewhere by Cabral and Hortacsu, 2010 on eBay. They have found that, “when a seller first receives negative feedback, his weekly sales rate drops from a positive 5% to a negative 8%; subsequent negative feedback ratings arrive 25% more rapidly than the first one and don't have nearly as much impact as the first one . . . a seller is more likely to exit the lower his reputation is; just before exiting, sellers receive more negative feedback than their lifetime average.“

4.2 Explanations linked to the rules of the website Bitcoin-OTC

A first “rule“ explanation lies in the computation mode of the reputation on the Bitcoin-OTC website. The reputation of an agent is a sum of all the scores that he/she has received since his/her arrival on the website. This tends to increase the number of positive scores and decrease the number of negative scores.

Indeed, this model of reputation computation favors the first “entering“ sellers in the website. The older sellers have the more positive reputations, and are thus more chosen for transactions. This is probably why we have observed that the nodes with high-degree in the rewarding layer are the most active. The effect is particularly strong, since, as outlined by Yaniv and Kleinberger, 2000, people tend to categorize possible choices in terms of good and bad agents, with good agents having the higher reputations, and bad having the average and lower reputations. Differing even more strongly good sellers from the others due to the reputation computation mode increases the probability to be satisfied by a transaction with them, and consequently the probability of rewards.

Also we can guess that the temporal speed of the increase of good reputations depends on the number of sellers considered as good on the market, the more concentrated is the market, the higher the temporal slope. Differently, the score trajectory keeps the same slope since most of the transactions correspond to new encounter valued +1 in case of satisfaction. This computation mode disfavors new sellers coming later on the website. On the contrary, it can favor users who have felt for a while since they can be distinguished at first sight from a recent seller. Then they have the opportunity to see their reputation rising again, and to become a controversial user.

Moreover, it has been shown that giving the internet user a vision of the whole story of sellers increases their credibility Dellarocas, 2006. It has also been pointed out that users prefer sellers having a lot of accumulated opinions than sellers scored few times Carbonell Carrasquilla, n.d. A reputation based on many accumulated opinions is based as less biased.

However, this computation mode for reputation is far from what has been observed off-line. Yaniv and Kleinberger, 2000 conclude from their experiments that "Reputation formation seems to be a heuristic process (such as confirmation bias, primacy effect, and negativity effect) that is potentially reactive toward negative information, subject to the effect of recent trials, and based on quick generalizations made on the basis of few data points". If we advocate for a more « natural » reputation, such a conclusion argues for another computation mode of the reputation stressing out the impact of recent events, especially the negative ones.

A second “rule“ explanation lies in the scale of the scores of the Bitcoin-OTC website. The scale is from -10 to +10 It is recommended by the website to put +1 as a reward for a first satisfying encounter. On the other hand, there is no recommendation for a dissatisfying experience. Thus the difference of amplitude between +1, and a possible first negative encounter scored -10, is such that the reputation is very sensitive to negative feedback. Moreover one negative feedback tend to increase the probability of another negative feedback as described earlier by the result of Cabral and Hortacsu, 2010. Such a vicious circle can slowly push the sellers exiting from the website since his/her reputation is getting down and down and it has less and less transactions. These exits also explain the lower frequency of punishments which can't continue forever, while rewards can. Overall, the asymmetry of the scale of scores is very important for the resilience of sellers who can be perceived as incompetent or cheaters for a while during few transactions.

It is recommended by the website to put +1 for a reward at the first encounter. However, we have shown that almost all the times, there is only one transactions between two same people. This can explain the slow slope of increase of the reputations even if a minority of people has not followed the recommendation.

Indeed, some users, probably relying on intermediary users, rate higher than 1 at their first transaction. This is a real peculiarity of social systems, which can't be easily governed by an a priori institution: social norms emerge from the interaction between users and are hardly predictable. This is particularly true when the behaviour to adopt is advised by trusted users as shown by the Theory of Reasoned Action Ajzen, 2001. This theory has pointed out how important is the subjective norm, the norm recommended by the important others, in the decision-making regarding the behaviour to adopt.

Our final “rule“ explanation relates to moderating or not moderating the exchange. The fact that rates can be given even if no monetary exchange happened (forum or discussion related issues are frequently mentioned in the comments though) can give rise to collective retaliation behaviors or massive misuse of rating system. This website has no "warranting principles", especially not the one consisting in checking if the writer of the review has bought the product Dellarocas, 2006. Thus, it can be attacked by trolls, but simultaneously defended by a coalition of observers.

Whitson *et al.*, 2015, from their study on norm maintenance, have pointed out that “The constant presence of third-party observers in organizational settings (e.g., managers, supervisors, coworkers, and subordinates) also makes them important rewarders and punishers. Data also suggest that observers can play a critical role in the development and mainte-

nance of norms of reciprocity.“

A last point which has not really being studied in this paper is the fact that this is possible to reward or to punish a user that a buyer has used as an advisor to choose his/her seller. This can also have important impact on the emergence and maintenance of norms and reputations.

Our work has stressed out first results and related issues that deserve more study. Especially this would be more interesting for future research to compare the rise and fall of reputation for different website in order to disentangle what comes from usual human behavior, from what comes from web design.

Acknowledgments

The authors thank Guillaume Deffuant for the useful discussions. This work has partly supported by the FuturICT2.0 Flagera project and by the CNRS-INFINITI project.

References

Adler, B. T. and L. De Alfaro (2007). “A content-driven reputation system for the Wikipedia”. In: *Proceedings of the 16th international conference on World Wide Web*. ACM. 261–270.

Ajzen, I. (2001). “Nature and operation of attitudes”. *Annual review of psychology*. 52(1): 27–58.

Antal, T., P. L. Krapivsky, and S. Redner (2006). “Social balance on networks: The dynamics of friendship and enmity”. *Physica D: Nonlinear Phenomena*. 224(1-2): 130–136.

Axelrod, R. (1984). “The evolution of cooperation Basic Books”. *New York*.

Bertazzi, I., S. Huet, G. Deffuant, and F. Gargiulo (2018). “The anatomy of a Web of Trust: the Bitcoin-OTC market”. In: *International Conference on Social Informatics*. Springer. 228–241.

Bonaccio, S. and R. S. Dalal (2006). “Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences”. *Organizational behavior and human decision processes*. 101(2): 127–151.

Bosu, A., C. S. Corley, D. Heaton, D. Chatterji, J. C. Carver, and N. A. Kraft (2013). “Building reputation in stackoverflow: an empirical investigation”. In: *2013 10th Working Conference on Mining Software Repositories (MSR)*. IEEE. 89–92.

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

Carbonell Carrasquilla, G. A. (n.d.). “Decisions based on ratings, reviews, and recommendations-the cognitive processing of online information”. *PhD thesis*.

Conte, R. and M. Paolucci (2002). *Reputation in artificial societies: Social beliefs for social order*. Vol. 6. Springer Science & Business Media.

Cuesta, J. A., C. Gracia-Lázaro, A. Ferrer, Y. Moreno, and A. Sánchez (2015). “Reputation drives cooperative behaviour and network formation in human groups”. *Scientific reports*. 5: 7843.

Dellarocas, C. (2006). “Reputation mechanisms. Forthcoming in Handbook on Economics and Information Systems.(T. Hendershott, ed.)” *Elsevier Publishing. J. Mundinger and J.-Y. Le Boudec. Systems and Beyond. In Proceedings of Inter-Perf*. 6: 6–23.

Dellarocas, C. and C. A. Wood (2008). “The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias”. *Management science*. 54(3): 460–476.

Guha, R., R. Kumar, P. Raghavan, and A. Tomkins (2004). “Propagation of trust and distrust”. In: *Proceedings of the 13th international conference on World Wide Web*. ACM. 403–412.

Heider, F. (1944). “Social perception and phenomenal causality.” *Psychological review*. 51(6): 358.

“https://bitcoin-otc.com” (n.d.). *last access: 13/05/2018*.

Kollock, P. *et al.* (1999). “The production of trust in online markets”. *Advances in group processes*. 16(1): 99–123.

Lauterbach, D., H. Truong, T. Shah, and L. Adamic (2009). “Surfing a web of trust: Reputation and reciprocity on couchsurfing. com”. In: *2009 International Conference on Computational Science and Engineering*. Vol. 4. IEEE. 346–353.

Leskovec, J., D. Huttenlocher, and J. Kleinberg (2010). “Signed networks in social media”. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM. 1361–1370.

Manzo, G. and D. Baldassarri (2015). “Heuristics, interactions, and status hierarchies: An agent-based model of deference exchange”. *Sociological Methods & Research*. 44(2): 329–387.

Merton, R. K. (1988). “The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property”. *isis*. 79(4): 606–623.

Movshovitz-Attias, D., Y. Movshovitz-Attias, P. Steenkiste, and C. Faloutsos (2013). “Analysis of the reputation system and user contributions on a question answering website: Stackoverflow”. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM. 886–893.

O’Connor, P. (2008). “User-generated content and travel: A case study on Tripadvisor. com”. *Information and communication technologies in tourism 2008*: 47–58.

Resnick, P., K. Kuwabara, R. Zeckhauser, and E. Friedman (2000). “Reputation systems”. *Communications of the ACM*. 43(12): 45–48.

Richardson, M., R. Agrawal, and P. Domingos (2003). “Trust management for the semantic web”. In: *International semantic Web conference*. Springer. 351–368.

Scellato, S., A. Noulas, R. Lambiotte, and C. Mascolo (2011). “Socio-spatial properties of online location-based social networks”. In: *Fifth international AAAI conference on weblogs and social media*.

Skowronski, J. J. and D. E. Carlston (1989). “Negativity and extremity biases in impression formation: A review of explanations.” *Psychological bulletin*. 105(1): 131.

Slovic, P. (1993). "Perceived risk, trust, and democracy". *Risk analysis*. 13(6): 675–682.

Thierer, A., C. Koopman, A. Hobson, and C. Kuiper (2015). "How the internet, the sharing economy, and reputational feedback mechanisms solve the lemons problem". *U. Miami L. Rev.* 70: 830.

Whitson, J. A., C. S. Wang, Y. H. M. See, W. E. Baker, and J. K. Murnighan (2015). "How, when, and why recipients and observers reward good deeds and punish bad deeds". *Organizational Behavior and Human Decision Processes*. 128: 84–95.

Yaniv, I. and E. Kleinberger (2000). "Advice taking in decision making: Egocentric discounting and reputation formation". *Organizational behavior and human decision processes*. 83(2): 260–281.

Yin, D., S. D. Bond, and H. Zhang (2010). "Are Bad Reviews Always Stronger than Good? Asymmetric Negativity Bias in the Formation of Online Consumer Trust." In: *ICIS*. 193.

Zeng, H., M. A. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness (2006). "Computing trust from revision history". *Tech. rep.* Stanford Univ Ca Knowledge Systems LAB.