University of Groningen

When is ignorance bliss? Disclosing true information and cascades of norm violation in networks

Maes, Michael; Opp, Karl-Dieter

Published in:
Social Networks

DOI:
[10.1016/j.socnet.2016.05.004](https://doi.org/10.1016/j.socnet.2016.05.004)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2016

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Maes, M., & Opp, K-D. (2016). When is ignorance bliss? Disclosing true information and cascades of norm violation in networks. *Social Networks*, 47, 116-129. <https://doi.org/10.1016/j.socnet.2016.05.004>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

When is ignorance bliss? Disclosing true information and cascades of norm violation in networks

Michael Mäs^{a,*}, Karl-Dieter Opp^{b,c}

^a University of Groningen, The Netherlands

^b University of Leipzig, Germany

^c University of Washington, USA

ARTICLE INFO

Article history:

Available online 23 June 2016

Keywords:

Social norms
Norm violation
Threshold model
Network clustering
Cascade
Nudging

ABSTRACT

It has been hypothesized that disclosing a population's true rate of norm violation increases norm-violating behavior. Withholding such information might, thus, prevent the attenuation of useful norms. Analyzing a classical threshold model with flexible thresholds, we show that disclosing the true rate of norm violation can spark cascades of norm violation but can also have the opposite effect, decreasing norm violation and strengthening norm acceptance. The direction of the cascade depends on the initial rate of norm violation. Furthermore, the disclosure effect depends on whether or not the rate of norm violation is disclosed repeatedly, the structure of the social network, and whether individuals' norm acceptance is inelastic or open to peer-influence.

© 2016 Elsevier B.V. All rights reserved.

False beliefs appear to be undesirable. For instance, false beliefs about a bank's allegedly limited solvency can lead to a bank run and induce the destabilization of the bank. On the other hand, generating false beliefs can be socially beneficial and morally acceptable. For example, a physician may downplay the severity of a patient's illness if she expects the patient to commit suicide. Generating this false belief will certainly be regarded as morally acceptable. Likewise, classical philosophers discuss the legitimacy of so-called 'noble lies', famously arguing that lying to the people may be justified when it preserves social order (Allen, 2008; Hoffmann et al., 2013).

In this paper we study the societal effects of specific kinds of false beliefs, viz. beliefs about the amount of norm violation in society. It happens time and again that we read news about unexpectedly frequent norm violations. An example is the large-scale sexual abuse of children by the clergy in 2009/2010, which was kept secret by church officials and took outsiders by surprise. Another example is the surprisingly high extent of plagiarism in German academia. After it became evident that Karl-Theodor zu Guttenberg, Germany's former minister of defense, faked his dissertation, numerous further instances of plagiarism were publicized. It was certainly unexpected, at least by non-academics, that there is such

a large extent of cheating among scholars. These examples have in common that many individuals were not aware of the high degree of norm violation and were surprised when credible information was published.

What are the consequences of disclosing the true rate of norm violation? Here, we study an answer to this question by Heinrich Popitz (1925–2002), who hypothesized that there is a *preventive effect of ignorance* (Popitz, 1968, 2006).¹ The underlying theoretical argument is simple. He argued that disclosing a population's true rate of unexpectedly frequent norm violation leads to more norm violation, and lower norm acceptance. Having learned that there is more norm violation than expected, Popitz explains, norm-complying individuals will decide to violate the norm. Furthermore, realizing that many members of the society violate the norm, individuals will decrease their norm acceptance, the extent to which complying with a norm is intrinsically valuable to an individual. This will aggregate to a collective decay of the norm. Popitz concluded that withholding information about the degree of norm violation in a population is socially desirable, as it prevents an increase in norm violation and preserves useful norms. There is

* Corresponding author at: Department of Sociology/ICS, University of Groningen, Grote Rozenstraat 31, 9712 TG, The Netherlands.
E-mail address: m.maes@rug.nl (M. Mäs).

¹ The small book from 1968 is part of a larger book from 2006. Quotations refer to the book from 2006. The German expression for "preventive effect of ignorance" (which is our translation) is "Präventivwirkung des Nichtwissens". Recent publications that address Popitz's theory are Diekmann et al. (2015), Groeber and Rauhut (2009), Rauhut (2013), and Opp (2011).

thus a “preventive effect of ignorance,” as Popitz puts it. Supporting Popitz’s theoretical argument, empirical studies found that norm violation increases when individuals observe deviant behavior of others (Keizer et al., 2008; Keizer et al., 2013; Keuschnigg and Wolbring, 2015; Thaler and Sunstein, 2009).

Popitz’s theoretical argument can be criticized for focusing only on those individuals who complied with the norm prior to the disclosure. If the true rate of norm violation is unexpectedly high, however, then there must be individuals who violated the norm before disclosure. How will these norm violators react to the disclosed information? According to Popitz’s own theoretical argument, these individuals might have violated the norm because they overestimated the rate of norm violation. Having learned that the true rate of norm violation is lower than they expected, some norm violators will refrain from deviating, an outcome that was also documented in empirical research (Cabinet-Office, 2012; Rauhut, 2013; Schultz et al., 2007; Thaler and Sunstein, 2009). In a recent field experiment (Cabinet-Office, 2012), for instance, official letters were sent to taxpayers calling on them to pay any tax debt. Some of the letters contained the true information that 9 out of 10 citizens pay their tax on time. This information increased the response rate from 67.5% in the control condition to up to 83%. Thus, disclosed information about deviance increased compliance and not norm violation, which is the exact opposite of what Popitz predicted.

This criticism suggests that, even though individuals might react to the disclosed information, the aggregate effects of disclosing the true rate of norm violation might be much weaker than Popitz predicted. In fact, recent empirical research supports this. A laboratory experiment conducted by Rauhut (2013) found strong individual reactions after the disclosure but no macro-effect. Rauhut found decreasing rates of lying amongst dishonest participants after they had been informed about the true rate of lying in the population. Amongst the honest participants, however, lying increased to a similar extent, such that lying remained unchanged on the aggregate level.

Here, we identify so far overlooked, structural conditions under which disclosing the true rate of norm violation leads to substantial changes in collective norm violation, even though individuals might react in opposite ways to the disclosed information. Contrary to the existing literature, which focused on disclosure effects on isolated individuals (Cabinet-Office, 2012; Schultz et al., 2007; Thaler and Sunstein, 2009), we study individuals integrated in networks and analyze the dynamics that emerge on the collective level. Existing research showed, for instance, that individuals react stronger when the disclosed information is accompanied by injunctive messages (Schultz et al., 2007; Thaler and Sunstein, 2009). These findings are interesting and important, but they focus on the reactions of isolated individuals, which neglects that actors who adjusted their behavior after having received information about norm violation might motivate their network contacts, who may not have reacted when the information was disclosed, to also adjust their behavior. These adjustments, in turn, might motivate third actors to reconsider their choices, and so on. Assume, as an example, a population of dog owners who have been informed about the true amount of dog fouling in their town. It is possible that most dog owners accurately estimated the rate of norm violation and, therefore, do not change their behavior. However, there might be a neighborhood where some dog owners had underestimated the rate of norm violation and now decide to no longer clean up behind their dogs. Observing the additional dog mess, also their neighbors might change their behavior, which in turn might motivate others to also violate the norm. Such *behavioral cascades* can profoundly change collective behavior over and above the disclosure effect on isolated individuals that existing studies documented.

Our aim is to develop new and testable hypotheses about the structural conditions under which disclosing the true rate of

norm violation sparks a cascade that aggregates to increased or decreased collective norm violation. To this end, we take as given that individuals adjust their behavior when they are informed about the true rate of norm violation and develop testable hypotheses about the conditions under which cascades are sparked off and norm violation spreads in the network. We study three structural factors, analyzing how changes in aggregate norm violation depend on (i) the degree of clustering in the network, (ii) the magnitude of the disclosed rate of norm violation, and on (iii) whether the true rate of norm violation is disclosed only once or repeatedly.

To this end, we analyze a probabilistic version of the classical threshold models (Granovetter, 1978; Meade and Islam, 2006; Oliver et al., 1985; Valente, 1996) which are standard in the study of diffusion in networks (Barash et al., 2012; Centola and Macy, 2007; Funk et al., 2010; Krassa, 1988; Meade and Islam, 2006; Rogers, 1995; Valente, 1996) and research on political mobilization (Kuran, 1989; Macy, 1991; Oliver et al., 1985; Schelling, 1978). In our model, the probability that an agent violates the norm depends on her beliefs about the behavior of others and her norm acceptance. The latter represents the degree to which complying is intrinsically valuable to the individual and is mathematically equivalent to thresholds in the classical threshold models. Like in the classical threshold models, our agents compare these thresholds with the behavior they observe in others. Before the disclosure, agents only observe the amount of norm violation amongst their network contacts. When the true rate of norm violation is disclosed, however, agents base their decisions on the rate of norm violation in the whole network. We show that this switch from local to global information can fundamentally change dynamics of norm violation and that the structure of the network influences the intensity of this disclosure effect.

Our model goes beyond standard threshold models in that we also study settings with *adaptive thresholds*. Inspired by Popitz’s informal reasoning, we include the assumption that thresholds are open to peer influence (Festinger et al., 1950; Wood, 2000), assuming that norm acceptance is weakened when agents observe more norm violation than they expected. Likewise, norm acceptance is intensified when agents observe high degrees of norm abiding behavior in others. We demonstrate that allowing thresholds to adapt in this form has profound effects on model dynamics and predictions concerning the consequence of disclosing the true rate of norm violation.

The new model is very general in that it can be applied to all binary decision problems with so-called positive externalities (Lopez-Pintado and Watts, 2008), cases where actors’ tendency to adopt a behavior is intensified when they observe it in others. Typical examples are cases where individuals seek to coordinate with their contacts. For instance, adopting a new communication technology (e.g. fax machine, Facebook) is more attractive when many use it. Also collective-good problems have positive externalities when the production function is convex. Soccer players, for instance, might increase own efforts in order to raise their team’s chances of winning when other team members also do their best. Norm violating behavior has positive externalities when individuals believe that norm violation is being punished and infer the chances of punishment from the observed amount of norm violation.

The remainder of this paper is organized as follows. In the next section, we formulate four research questions about the structural conditions of disclosure effects and demonstrate that a formal analysis is needed to validly derive answers to these questions. Subsequently, we describe the formal model that we analyzed to answer the research question and present our analyses. Finally, we summarize our results in a set of new hypotheses, and point to future work.

1. Four research questions

Deriving hypotheses about the collective consequences of disclosing the true rate of norm violation is intricate. For instance, will more or fewer dog owners remove their pet's feces when the disclosed rate of dog fouling is 70 percent rather than 90 percent? On the one hand, intuition suggests that the effect on those who clean up behind their dog is stronger when the magnitude of the disclosed rate is high, which suggests a stronger effect in the 90-percent case. On the other hand, when the norm violation amounts to 90 percent, then there are relatively few actors left to turn into norm violators. Furthermore, there must be a reason why the rate of norm violation is so high in the 90-percent case. Very likely some of the norm violating dog owners overestimated the rate of norm violation, falsely believing that the rate of norm violation even exceeded 90 percent. Having learned that the true rate amounts to only 90 percent, at least some of the overestimators will start removing their dog's waste, which might trigger off a cascade in the opposite direction. As there are so many norm violators, such a cascade has the potential to substantially decrease norm violation. At least in the 90 percent case it is, therefore, not clear whether or not the number of those who underestimated the rate of norm violation and turn into violators exceeds the number of those who overestimated the rate and turn into norm followers. In sum, one can come up with conflicting intuitions about the effects of the magnitude of the disclosed rate, which leads us to our first research question.

Q1: When the true rate of norm violation is disclosed, does the magnitude of the disclosed rate of norm violation influence whether norm violation increases, decreases or remain constant?

Likewise, it is difficult to intuitively derive a hypothesis about the effects of network structure on the disclosure effect. On the one hand, network structure seems to play an important role. Assume, for instance, a population where every node is able to observe the behavior of 10 other randomly chosen nodes. In this random network, actors base their estimates of the rate of norm violation on random samples, which should lead to rather precise estimates in most cases. As a consequence, the true rate will not come as a great surprise for most agents and the disclosure effect on behavior should be limited. Dog owners, in contrast, tend to walk their dog in the neighborhood where they live. Even though dog owners might not directly observe each other's behavior, they will have a good estimate of the rate of norm violation in their neighborhood, knowing the number of local dogs and observing the amount of dog waste on the street. This creates a clustered network where two nodes whose behavior is observed by a third node are very likely also observing each other's behavior. In this case, it is plausible that some of the network clusters (neighborhoods) will happen to consist of many norm violators and that the local rate of norm violation is high. Other clusters might happen to consist of many norm-abiding actors, which results in a low local rate of norm violation. Therefore, if this clustered population is informed about the global rate of norm violation, many actors will be surprised and adjust their behavior. Furthermore, network clustering makes cascades more likely, because clustered networks consist by definition of many closed triplets. If actor A in a closed triplet of norm followers (A, B, C) turns into a violator, then there is a certain chance that actor B also starts violating the norm. The crucial characteristic of a closed triplet is that now actor C observes two additional norm violators, which strongly increases the chances that also C will start violating the norm.

On the other hand, if there are clusters of norm abiding actors and if these actors are surprised by the magnitude of the disclosed rate, then the network must also consist of clusters of norm violating individuals. This, in turn, implies that the global rate of norm

violation will be neither very low nor very high. As a consequence, there will be both many actors who turn into norm violators and actors who start abiding by the norm. Contrary to the first intuitive prediction, this suggests that the total change of the rate of norm violation is limited also in clustered networks. Our second research question asks which of the two conflicting intuitions is correct.

Q2: Does the degree of network clustering influence the degree to which norm violation is affected by disclosing the rate of norm violation?

Deriving hypotheses is even more difficult when one considers Popitz's assumption that the degree to which individuals are intrinsically motivated to abide by the norm, i.e. their norm *acceptance*, is affected by the amount of norm violation that they observe. In particular, Popitz argued that disclosing the true rate of norm violation leads to norm decay because learning that many violate the norm will decrease the intrinsic motivation of individuals to follow the norm. Obviously, this assumption is not always plausible. For instance, most individuals will not change their normative beliefs about extreme norm violation such as child abuse, or murder when they learn that this behavior is more frequent than they thought. However, dog owners who are convinced that one should remove dog waste might be frustrated when they observe that the majority of dog owners violate the norm. Likewise, dog owners who observe that many others clean up behind their dog might feel morally obliged to do the same, a process that is called norm internalization. Another case where norm acceptance appears to be relatively flexible is fraudulent behavior in academia. A researcher who believes that the majority of her colleagues in the field cite the work of friends without scientific reason and only seek to form citation cartels (Franck, 1999) might change her normative beliefs, feeling that she has to adopt this behavior in order to not be disadvantaged. However, in a field where such behavior is very rare, researchers will likely develop a strong intrinsic motivation to abide by this norm.

On the one hand, it appears plausible that disclosure effects are intensified when norm acceptance is flexible. For instance, disclosing that very few dog owners remove dog waste will not only affect the behavior of those who underestimated the amount of norm violation. In addition, dog owners who had an accurate estimate but cleaned up behind their pet because they were intrinsically highly motivated might decide to also violate the norm because their norm acceptance decreased. More importantly, however, adjustments in norm acceptance have the potential to reinforce cascades, as they generate a self-reinforcing feedback process. For instance, observing the increase in dog fouling after the disclosure, dog owners who used to be highly motivated to abide by the norm will decrease their norm acceptance, which will lead some of them to also violate the norm. The resulting additional increase in dog waste will lead to further decrease norm acceptance, which will again increase norm violation and result in declining norm acceptance and so on.

On the other hand, when norm acceptance is open to peer influence, why haven't actors' intrinsic motivations already converged before the disclosure? In fact, formal models of social-influence dynamics in networks show that influence generates convergence cascades that lead networks toward perfect homogeneity, unless the network consists of perfectly disconnected subgraphs (Abelson, 1964). This suggests that in settings where norm acceptance is open to peer influence, actors have similar norm acceptance, and, therefore, violate the norm with the same probability. As a consequence, disclosing the true rate of norm violation will not come as a big surprise to most actors and will have a limited effect. These conflicting intuitions led to our third research question.

Q3: Is the change in collective norm violation stronger when norm acceptance is open to peer influence?

Finally, we sought to derive predictions about the effects of repeated disclosure. Assume again that a population of dog owners is informed about the amount of norm violation and that some dog owners decide to not remove their dog's feces any longer. Observing more dog waste, also owners who were not directly affected by the disclosure might change their behavior. Even though such behavioral cascades will emerge, there is a good chance that many of them are limited and level off over time. What will happen, however, when officials commission regular reassessments of norm violation and always publish updated results? One possible scenario is that every time the rate of norm violation is disclosed new cascades are sparked off, which creates a self-reinforcing process that ends when all dog owners violate the norm. Alternatively, it seems possible that the number of new norm violators decreases after each round of disclosure, which would imply that the disclosure effect diminishes over time. Our fourth research question asks which of the two scenarios follows from the formal model.

Q4: How is the long-term rate of norm violation affected when the rate of norm violation is disclosed repeatedly?

2. The formal model

We derived answers to the four research questions from a probabilistic threshold model. Like in the classical threshold models, we assumed a population of N individuals i , called agents. At each point in time t , every agent i either engages in norm violating behavior ($v_{i,t} = 1$) or not ($v_{i,t} = 0$). Furthermore, a pair of agents can be socially connected by a network tie. Connected agents can observe each other's behavior $v_{i,t}$ but have no direct knowledge of any further individual characteristics. All network ties are undirected (i.e. if agents i is connected to j then j is linked to i , too) and are not weighted.

Whether an agent violates the norm or not depends on the probability of norm violation, which is represented by $P_{i,t}(v_{i,t} = 1)$. Like all probabilities, $P_{i,t}(v_{i,t} = 1)$ can adopt any value between zero and one. A value of zero means that the agent i will not violate the norm ($v_{i,t} = 0$). The higher the value of $P_{i,t}(v_{i,t} = 1)$ the more likely i violates the norm ($v_{i,t} = 1$).

The probability of norm violation is calculated based on Eq. (1), which we adopted from existing threshold models (e.g. Macy, 1991). It models the probability of norm violation at time point $t + 1$ as an S-shaped (logistic) function of the gap between agent i 's estimated rate $r_{i,t}$ of norm violation and i 's norm acceptance $a_{i,t}$. The estimated rate $r_{i,t}$ and norm acceptance $a_{i,t}$ adopt values between zero and one. Eq. (1), therefore, implies that the probability of norm violation is small when agent i believes that the population's rate of norm violation is low and when her own norm acceptance is high. The probability of norm violation reaches .5 when $r_{i,t} = a_{i,t}$. For instance, an agent with a very high norm acceptance ($a_{i,t} = .9$) will violate the norm with probability 50% when she believes norm violation in others to be very high too ($r_{i,t} = .9$).

$$P_{i,t+1}(v_{i,t} = 1) = \frac{1}{1 + e^{-\delta(r_{i,t} - a_{i,t})}} \quad (1)$$

The slope parameter δ ($\delta > 0$) determines the shape of the logistic function and allows one to manipulate the amount of randomness in agents' behavior (McFadden, 1973). If $\delta = 0$, for instance, then the decision to violate is completely random ($P_{i,t}(v_{i,t} = 1) = 0.5$ for all $r_{i,t}$ and $a_{i,t}$). For positive value of δ Eq. (1) implies that the probability of norm violation increasingly depends on ($r_{i,t} - a_{i,t}$), rather than randomness. As the term ($r_{i,t} - a_{i,t}$) can vary only between -1 and 1

in our model, the logit-function resembles a linear function when δ adopts small values. However, it has its characteristic S-shape when δ adopts values of 5 or higher. When δ equals infinity, the logistic function is a deterministic step function, which implies a probability of norm violation of zero when $r_{i,t} < a_{i,t}$ and a probability of one when $r_{i,t} > a_{i,t}$.

So far, the model is a replication of existing threshold models. We added, however, two model ingredients. First, Eq. (2) specifies how each agent forms her estimate $r_{i,t}$ of the rate of norm violation. It implies that agents can base their estimate on the local rate of norm violation amongst their n_i network contacts and/or the global rate of norm violation in the entire population. Parameter B_t allows manipulating the relative impact of the two rates. Most importantly, when $B_t = 0$ agents are not informed about the true rate of norm violation and base their estimate $r_{i,t}$ on the behavior that they observed amongst their network contacts. If $B_t = 1$, agents' estimated rate of norm violation is equal to the true rate in the whole population at the previous point in time. Varying parameter B_t over time allowed us to study dynamics in populations where agents first estimated the rate of norm violation based on the behavior of their network contacts ($B_t = 0$) and were then informed about the true rate of norm violation ($B_t = 1$).

$$r_{i,t} = \frac{1 - B_t}{n_i} \cdot \sum_{j=1}^{n_i} v_{j,t-1} + \frac{B_t}{N} \cdot \sum_{j=1}^N v_{j,t-1} \quad (2)$$

The second new ingredient is Eq. (3), which includes Popitz's assumption that norm acceptance is flexible and open to peer influence. Eq. (3) assumes that agent i 's norm acceptance at time point t is a function of her previous norm acceptance and her estimate of the rate of norm violation amongst the other agents. Implemented in a similar fashion as in existing models of social influence in networks (Abelson, 1964; DeGroot, 1974; Friedkin and Johnsen, 2011; Hegselmann and Krause, 2002), the latter assumes that norm acceptance decreases when agents believe that global norm violation is high ($r_{i,t} > a_{i,t}$) and increases when agents perceive infrequent norm violation amongst others ($r_{i,t} < a_{i,t}$). Only when the estimated rate of norm violation equals the current norm acceptance ($r_{i,t} = a_{i,t}$), norm acceptance remains unchanged. Parameter γ weights the relative impact of peer influence ($0 \leq \gamma \leq 1$). The larger the value assigned to γ , the stronger is peer influence.

$$a_{i,t} = (1 - \gamma)a_{i,t-1} + \gamma \cdot (1 - r_{i,t}). \quad (3)$$

The assumption that thresholds adapt has been studied before. Most prominently, Macy (1991) studied learning effects, assuming that thresholds adapt in reaction to positively or negatively experienced outcomes in the past. Contrary to this, thresholds in our model adapt to the perceived behavior of others, implementing peer influence rather than learning.

2.1. Equilibrium analysis

Strictly speaking, the model will never reach a state of perfect equilibrium where all agents will stick to their behavior $v_{i,t}$ for ever. This follows from the probabilistic nature of the model and the assumption that norm acceptance and the estimated rate of norm violation range between zero and one. The latter implies that $P_{i,t}(v_{i,t} = 1)$ will never be exactly zero or one, unless the slope parameter δ adopts infinity (see Eq. (1)). In other words, even when agents have a maximal norm acceptance and perceive zero norm violation, there is always a tiny but non-zero probability that the agent will violate the norm. However, for infinitely big populations the model can reach states where the collective rate of norm violation remains constant, even though individual agents might still change their behavior.

To identify the equilibria of the model and study whether violation rates increase or decrease after the true rate of norm violation has been disclosed ($B_t = 1$), we assume that all agents simultaneously and immediately ($\gamma = 1$) update their norm acceptance $a_{i,t}$, which simplifies the whole model into one equation.

$$P_{i,t+1}(v_{i,t} = 1) = \frac{1}{1 + e^{-\delta(r_{i,t} - (1 - r_{i,t}))}} \quad (1a)$$

The left panel of Fig. 1 plots Eq. (1a) for $\delta = \infty$. If the plotted line lies above the diagonal, then agent i will violate with an increased probability. If it lies below, the probability of norm violation will decrease. The three dots identify all intersections of the function and the diagonal (dashed line), showing perceived rates of norm violation that result in identical probabilities of norm violation. Thus, if all agents perceive these rates, then the population's updated rate of norm violation will adopt the same value and equilibrium is reached (in an infinitely big population).

There are three equilibria when $\delta = \infty$. First, if all agents violate the norm, then all agents will stick to this behavior also in the next round. Second, if the rate of norm violation is zero, it will remain zero. These two rest points are stable in that even when agents happen to deviate from the equilibrium behavior, the system will always return to the same equilibrium. Third, when the rate of norm violation is exactly 0.5, agents will violate the norm with a probability of 50 percent. This equilibrium, however, is very fragile in finite populations, because random fluctuations will result in deviation rates smaller or bigger than 0.5, which will, in turn, immediately move the system into one of the other two equilibria.

When parameter δ adopts values smaller than infinity, two cases need to be distinguished. First, when δ adopts positive values below 2, as illustrated in the right panel of Fig. 1, then the system will always move toward a state where all agents violate with a probability of 0.5. That is, when the rate of norm violation is below 0.5, agents will increase norm violation (line is above the diagonal) and they will decrease their violation probabilities when the perceived violation rate exceeds 0.5 (line is below the diagonal). Parameter values of $\delta < 2$, however, are an unrealistically weak implementation of the model assumptions in that agents' behavior is too much driven by randomness rather than the model assumptions. Fig. 1 shows, for instance, that $\delta = 1$ implies that even a person with maximal norm acceptance who thinks that nobody violates the norm will violate the norm with a probability of 27 percent. In line with earlier studies (Macy, 1991), we therefore focused on the second case ($\delta > 2$).

The second and more relevant case obtains when δ exceeds 2, which is illustrated in the center panel of Fig. 1. Here, the state with a rate of norm violation of 0.5 is also an equilibrium, but random fluctuations, which will always occur in finite populations, will lead the system away from this state. If fluctuations happen to lead to a violation rate above 0.5, agents will further increase the probability of norm violation (line is above the diagonal). Otherwise, deviation rates will decline (line is below the diagonal). However, these processes do not continue until all agents either obey the norm or violate it. Contrary to the case with $\delta = \infty$, the other two equilibria are always characterized by some variation in agents' behavior. This results from the logit-function, which approaches probabilities of zero and one but never reaches these values. However, the panel for $\delta = 5$ illustrates that for larger δ the two extreme equilibria imply violation rates close to 0 or 1.

2.2. Simulation of the dynamics

To study the model's answers to our four research questions, we had to resort to computational agent-based modeling (Epstein, 2007; Macy and Flache, 2009; Macy and Willer, 2002), a rigorous formal modeling technique that, in contrast to analytical

techniques, imposes very few restrictions on the choice of model assumptions. This was particularly useful for the purpose of our study, as we sought to explore the implications of Popitz's assumption that norm acceptance is open to peer influence (Research Question 3), which considerably complicates model dynamics and analyses. In addition, this less restrictive modeling approach allowed us to explore alternative interpretations of Popitz's theory. For instance, we studied alternative scenarios of how the degree of norm violation is disclosed (Research Question 4) and different network structures (Research Question 2).

The simulation program implements that at the outset of the interaction process norm acceptance is distributed at random amongst the N agents. To be more precise, the initial norm acceptance ($a_{i,0}$) of each agent is drawn from a Beta distribution (Beta (α, β)), a probability distribution that is restricted to the interval $[0,1]$. The parameters α and β of the Beta distribution influence the shape of the resulting distribution of norm acceptance. In our simulation experiments, we fixed parameter β at 1 but experimented with different values of α . An $\alpha = 1$ generates a uniform distribution of norm acceptance in the population. An α below 1 results in a positively skewed distribution. Values above 1 lead to a negatively skewed distribution of norm acceptance. Agents' initial probability of norm violation ($P_{i,0}(v_{i,t} = 1)$) was assumed to depend only on their initial norm acceptance. The more an agent accepts the norm, the more likely he or she is to comply with the norm. Formally, $P_{i,0}(v_{i,t} = 1) = 1 - a_{i,0}$. The initial behavior ($v_{i,t}$) of the agents follows directly from the probability of norm violation. For instance, if an agent i has been assigned a norm acceptance of $a_{i,0} = 0.3$, then his initial probability of violating the norm equals 0.7. This, in turn, implies that there is a thirty percent chance that this agent violates the norm at time point zero.

Model dynamics were broken down to a series of discrete simulation events t . At each event, the behavior of all agents is updated synchronously. That is, all agents first update their perceived rate of norm violation (Eq. (2)) and their norm acceptance (Eq. (3)). Next, all agents update their probability of norm violation (Eq. (1)). Replicating all analyses with a model that uses asynchronous updating lead to virtually identical results, which is in line with Lopez-Pintado and Watts's (2008) analytical finding that the outcomes of threshold models are unaffected by the choice of the updating rule when so-called "positive externalities" are modeled as in the present study.

2.3. The simulation experiment

To study the answers that our model gives to the four research questions, we conducted a computer-simulation experiment. In this experiment, we studied networks of 100 agents, who had six contacts each. The dependent variable in our analyses was the change in the rate of norm violation that results from the disclosure of the true rate of norm violation.

In all simulation runs, the disclosure was implemented with the following schedule. First, we studied model dynamics for 100 simulation periods where all agents estimated the rate of norm violation only based on the behavior of their network contacts ($B_t = 0$ for $t < 100$). It turned out that after 100 periods the populations had reached a state where the rate of norm violation remained at least temporarily within a small interval. At time step 100, however, the true rate of norm violation was disclosed for up to 100 events. Technically, we set the value of parameter B_{100} to one, which meant that each agent's estimate of the rate of norm violation equaled the true rate at time point $t = 99$. Finally, simulations were continued for another 100 periods with $B_t = 0$ to measure the degree to which disclosing of the true rate of norm violation affected norm violating behavior also in the long run.

Fig. 1. Equilibrium analyses ($\gamma = 1$, $B_t = 1$).

The independent variable of Research Question 1 is the magnitude of the rate of norm violation at the moment when it is disclosed. It was not necessary to experimentally manipulate this variable, as it could be measured at $t = 99$.

Research Question 2 asks whether the degree of network clustering affects disclosure dynamics. To manipulate network clustering, we applied a standard method that was proposed by Maslov and Sneppen (2002). First, we created symmetric circle networks where agents are connected to the three closest circle neighbors to the right and to the left. Second, the program randomly rewired R pairs of network relationships according to the Maslov–Sneppen-rewiring algorithm. This algorithm randomly picks two pairs of connected agents (e.g. agents 1 and 2; and 8 and 9) that are not directly connected so far (thus, neither 1 nor 2 is connected to 8 or 9). Next, the two network relationships are dissolved and new relationships between the initially unconnected agents are created (e.g. between 1 and 9 as well as 2 and 8). The great advantage of this rewiring method is that it manipulates the degree of clustering without changing other crucial network characteristics such as the number of network ties or the degree-distribution.

We varied the number R of rewiring steps between 0 and 1000 ($R \in [0, 5, 10, 20, 30, 40, 50, 100, 1000]$). If $R = 0$, the network remains a perfectly symmetric circle. Under this condition, connected agents tend to form triplets such that the network contacts of an agent tend to be connected too. This is the core characteristic of highly clustered networks. In addition, the network has a high average shortest path length, which means that it requires many steps through the network to create a connection between any pair of agents. For small nonzero values of R ($R = 5$ and $R = 10$), the network remains highly clustered but shortest-path length decreases substantially, creating so-called “small worlds” (Watts and Strogatz, 1998). For higher values of R , the degree of clustering and mean-shortest path length decrease until a random network ($R = 1000$) with low clustering and a very small shortest-path length obtains.

To study Research Question 3, we compared model predictions in settings where norm acceptance is inflexible ($\gamma = 0$) to settings where norm acceptance is open to peer influence as implemented in Eq. (3). For the latter condition, we assumed that $\gamma = 0.2$, which implements that social influence changes norm acceptance rather slowly. We explored model dynamics under alternative values of γ , but did not find substantial differences. Therefore, we chose this relatively weak implementation of social influence.

To answer Research Question 4, we compared scenarios where the true rate of norm violation is disclosed only once or repeatedly. There were five different disclosure scenarios. First, in the control condition the true rate of norm violation is never disclosed ($B_t = 0$ for all t). In the second scenario, the true rate of norm violation at time point 99 is disclosed at simulation event $t = 100$ ($B_{100} = 1$). Subsequently, agents base their

estimate of the rate of norm violation again on the behavior of their neighbors ($B_t = 0$ for $t > 100$). We call this scenario *single disclosure with no memory*. The third scenario captures the situation where the rate of norm violation is published once but individuals base their decision to violate the norm or not on this value for a longer period. We implemented that from time point 100 to 199 agents base their decision on the actual rate of norm violation at time point 99 ($B_t = 1$ for $100 \leq t \leq 199$). This scenario is denoted as *single disclosure with long memory*. The fourth scenario resembles the third, but we included that the impact of the disclosed information decreases linearly over time. We call this scenario *single disclosure with decreasing memory*. Formally, we implemented

$$B_t = \frac{200 - t}{100} \quad \text{if } 100 \leq t \leq 200. \quad (4)$$

Finally, we included an experimental condition where the rate of norm violation is disclosed repeatedly and where the disclosed rate is updated at every point in time ($B_t = 1$ for $100 \leq t \leq 199$). This models a setting where individuals are informed regularly about the current rate of norm violation. We call this scenario the *repeated disclosure scenario*.

As a sensitivity test, we also studied the core control parameter of the formal model, δ , testing to which degree model implications change when random deviations from the behavioral patterns that our model presumes are taken into account. Studying a similar parameter range as existing contributions to literature on threshold models (Macy, 1991), we included conditions where δ either adopted a value of five (a lot of randomness), ten (moderate degree of randomness), or twenty (very little randomness). Furthermore, we tested whether our results are sensitive to the initial distribution of norm acceptance, varying parameter α between 1 and 2 in steps of 0.25. If $\alpha = 1$, then the initial norm acceptance is uniformly distributed (average = 0.5; sd = 0.20). The higher the value assigned to α , however, the more negatively skewed the distribution becomes. For instance, an α of two leads to an average norm acceptance of only 0.02 and a standard deviation of 0.02.

In total, the simulation experiment comprised 1350 ($= 9 \times 5 \times 2 \times 3 \times 5$) experimental conditions. We performed 100 independent replications for each of the conditions, leading to a total of 135,000 simulation runs.

3. Results

We report the outcomes of the simulation experiment in two steps, summarizing first the results when norm acceptance was fixed ($\gamma = 0$) and second the results when norm acceptance was flexible ($\gamma = 0.2$).

Fig. 2. Development of the distribution of rate of norm violation in 500 replications per disclosure scenario with fixed norm acceptance.

3.1. Fixed norm acceptance

Fig. 2 illustrates how disclosing the true rate of norm violation affects subsequent norm violation. To generate it, we measured the development of the rate of norm violation in 500 independent simulation runs for each of the five disclosure scenarios. We focused on one representative condition of the simulation experiment, imposing that norm acceptance is uniformly distributed ($\alpha = 1$), moderate amounts of noise ($\delta = 10$), and a clustered network ($R = 5$). Fig. 2 depicts the development of the distribution of the rate of norm violation in the 500 simulation runs. The white area shows the complete value range of the rate of norm violation. The light gray area represents the inter-quartile range and the dark gray area shows the value range of those 25 percent of the runs that are in the center of the distribution. The gray line in the center of the panels shows the average rate of norm violation in the 500 runs.

Fig. 2 shows that the rate of norm violation fluctuates over time even before the true rate norm violation has been disclosed. These fluctuations result from the probabilistic implementation of agents' decision rule ($\delta = 10$). In addition, there seems to be no significant difference between the conditions without any disclosure and the

disclosure scenarios with only a single disclosure. The bottom panel in Fig. 2, in contrast, shows that there is a strong effect when an updated rate of norm violation is disclosed repeatedly.

There are three interesting observations concerning the repeated disclosure scenario. First, disclosing the rate of norm violation either led to an increase or a decrease in norm violation. Both dynamics were observed equally often (see the average). This result contradicts Popitz's prediction that disclosure will always raise norm violation and is in line with the empirical findings from Rauhut's experiment (2013). Second, even though the repeated disclosure had a strong impact, it never led to situations where either all agents violated the norm or all agents complied with the norm. Third, when agents stopped making decisions based on the disclosed rate of norm violation ($t > 199$), the rate returned to the same level as before the disclosure, showing that the disclosure did not have long-term effects.

Providing statistical tests, Fig. 3 shows average differences in the rate of norm violation just before the disclosure ($t = 99$) and at four different points in time after the disclosure ($t = 100$, $t = 150$, $t = 200$, $t = 250$). As Fig. 2 showed that both an increase and a decrease in norm violation is possible, we report in Fig. 3 the absolute change in norm violation. These analyses are based on all simulation runs of the experiment where norm acceptance was fixed ($\gamma = 0$) and uniformly distributed at the outset ($\alpha = 1$).

The bar at the very left of Fig. 3 shows that in the control condition (no disclosure), the rate of norm violation changed only slightly from $t = 99$ to $t = 100$. In the other disclosure scenarios the immediate change in norm violation was significantly stronger. However, the difference is rather small, which is very surprising. At $t = 100$, changes in norm violation result only from agents' direct reaction to the disclosure without any reactions to others changes and resulting cascades. The fact that this effect is limited demonstrates the importance of cascades for explaining changes in norm violation.

The bar for $t = 150$ in the no-disclosure condition shows that there was a significant increase in the average absolute change in norm violation compared to $t = 99$ also without any disclosure of the rate of norm violation. This effect is caused by the random fluctuations that result from the probabilistic implementation of individuals' behavior, which always generates some change. As the figure shows absolute changes, these fluctuations are not averaged out.

Strikingly, the change in norm violation between $t = 99$ and $t = 150$ was significantly weaker in the conditions with single disclosure and long or decreasing memory than in the no-disclosure condition. This totally unexpected result obtained because under these conditions agents base their decision always on the same rate of norm violation. As a consequence, there are relatively small behavioral fluctuations only. When there is no disclosure, in contrast, agents respond to the behavior of their neighbors. In this process it is possible that random fluctuations in the behavior of a few agents can lead to behavioral changes of their neighbors, which reinforces the random fluctuations and can lead to a relatively strong temporal increase or decrease in the rate of norm violation. These small cascades are excluded when agents base their estimate of the rate of norm violation on the disclosed value and, thus, neglect recent changes in norm violation amongst their contacts.

Most importantly, Fig. 3 shows that the disclosure of the actual rate of norm violation had a substantial impact on norm violation when the disclosure was repeated. Under this condition, a self-reinforcing process unfolds where, e.g. a high rate of norm violation is disclosed and more agents violate the norm. When the increased rate of norm violation is disclosed, agent's probability of norm violation rises further, which leads to a further increase in the disclosed rate of norm violation, and so on. This self-reinforcing process, however, does not lead agents to always deviate, as norm

Fig. 3. Results of the simulation experiment with fixed norm acceptance (based on all runs of the experiment with $\alpha = 1$), error bars show 95 percent confidence intervals.

acceptance is uniformly distributed and fixed. According to Eq. (1), this implies that the probability of norm violation is for most agents much smaller than 1, independent of how much norm violation they perceive. What is more, even though there are strong short-term effects in the repeated disclosure scenario, there are no long-term effects. Already a single period after the final disclosure ($t=200$) the difference in norm violation compared to $t=99$ had decreased, as Fig. 3 shows. At $t=250$, the rate of norm violation adopted similar values as before the disclosure.

Did the magnitude of the disclosed rate of norm violation affect whether norm violation increased or decreased (Research Question 1)? Fig. 4 depicts how the disclosed rate of norm violation ($t=99$) affected the change in norm violation between $t=99$ and $t=150$, showing a scatterplot. In addition, a local polynomial smooth line shows whether we observed more often an increase in norm violation (line is above zero) or a decrease (line is below zero). As the rate of norm violation cannot fall below zero and cannot rise above one, we also included two straight lines that identify all logically possible combinations of the dependent and the independent variable.

In the simulation runs without disclosure, the rate of norm violation at $t=99$ hardly predicts the change in norm violation. The polynomial-fit line shows a weak effect, which obtains because the rate of norm violation can change only to a limited degree when it adopts high or low values already before the disclosure. In the scenario with a single disclosure and no memory, there is a very similar pattern as in the simulations without disclosure. This replicates our earlier observation that this disclosure scenario does not affect the dynamics.

Fig. 4, however, shows the opposite pattern for the scenario with a single disclosure and decreasing memory as well as the scenario with single disclosure and long memory. It turned out that the rate of norm violation tends to increase when it adopts values above 0.5 before the disclosure and vice versa. This pattern differs from the pattern observed under the no-disclosure condition, demonstrating that these forms of disclosure did affect the dynamics. Nevertheless, both disclosure

scenarios generated very limited changes in the rate of norm violation (see Fig. 3).

For the condition with repeated disclosure, however, Fig. 4 shows that the magnitude of the disclosed rate of norm violation had a strong impact on whether the described self-reinforced dynamics resulting from repeated disclosure led to increasing or decreasing norm violation. When, prior to the first disclosure, fewer than half of the agents violated the norm, then norm violation decreased. In contrast, norm violation tended to increase when the majority of the agents violated the norm already before the disclosure.

Finally, we analyzed whether or not the model predicts effects of the network structure (Research Question 2). It turned out that the degree of network clustering had no effects on the strength of the disclosure effect in the three scenarios with a single disclosure, which results from the very limited disclosure effects in these scenarios. However, when the true rate of norm violation was disclosed repeatedly, the change in norm violation after the disclosure was significantly stronger when the network was clustered. This supports our intuitive prediction that network clustering fosters the formation of subsets of connected agents that either have relatively high or low local norm violation. Members of such subsets have incorrect estimates of the global rate of norm violation, which leads to strong changes in the probability of norm violation after the true rate of norm violation has been disclosed.

3.2. Flexible norm acceptance

The model generates substantially more complex dynamics and different predictions when norm acceptance is flexible ($\gamma=0.2$). For illustration, Fig. 5 shows the dynamics of norm violation and norm acceptance in an ideal-typical simulation run where a highly clustered network ($R=0$) was repeatedly informed about the true rate of norm violation. The graph at the bottom of Fig. 5 shows the development of the rate of norm violation. In the gray boxes, we show network graphs where circles represent agents who violate the norm. Agents who comply with the norm are shown with

Fig. 4. Direction of the change in the rate of norm violation as a function of the rate of norm violation just before disclosure when norm acceptance is fixed (based on all runs of the experiment with $\alpha = 1$ and $\delta = 10$).

stars. The color of a network node informs about the agent's norm acceptance. We used shades of red where a very light red identifies agents with a high norm acceptance ($a_{i,t}$) and a dark red shows agents with a low norm acceptance. The extremes of this color scale are white (maximal norm acceptance) and black (minimal norm acceptance). The lower part of the five boxes contains histograms

of the distribution of norm acceptance at the respective points in time.

At the outset of the ideal-typical simulation ($t=0$), norm acceptance was uniformly distributed with an initial rate of norm violation of 0.51 (see the left box of Fig. 5). Norm acceptance and behavior was randomly distributed in the network. As agents' norm acceptance was influenced by the behavior of their network contacts, however, the distribution of norm acceptance changed dramatically. The network graph for Period $t=99$ shows that homogenous subgroups of agent with either maximal or minimal norm acceptance formed. As the subgroup of norm violators consisted of more agents, norm violation increased and then fluctuated between rates of .6 and .7. The bottom graph shows that this configuration was relatively stable. However, dynamics had not reached a state of equilibrium. Those agents that connected subgroups changed their behavior frequently. As a consequence, subgroups temporarily grew or shrunk in size. As this process resulted only from random fluctuations, however, it would have taken very long until either all agents would have adopted a minimal or a maximal probability of norm violation. At Period $t=100$, agents based their decision for the first time on the true rate of norm violation. In the subsequent periods, norm violation increased and reached a maximal level of one for the first time by $t=112$.

The following dynamics is responsible for the strong increase in norm violation after the disclosure. At $t=99$, most of the agents had a very incorrect estimate of the rate of norm violation in the system, either very much underestimating or very much overestimating. However, since there were more norm violators than norm complying agents, the estimates of norm violators were closer to the true value. As a consequence, norm violators adjusted their norm violation to a smaller degree than those agents who complied with the norm before the disclosure. This, in turn, implied that these agents' decrease in the probability of norm violation was smaller than the respective increase of norm complying agents. This aggregated to an increasing rate of norm violation. In addition, the disclosure also had an effect on agents' norm acceptance, which is only possible when norm acceptance is flexible. As the disclosed rate of norm violation was greater than 0.5, norm acceptance of those agents who followed the norm before disclosure decreased substantially. This decrease was stronger than the increase in norm acceptance in those agents who violated the norm already before the disclosure, leading to an additional decrease in norm acceptance and a further increase in norm violation. In the subsequent periods, agents made similar adjustments, generating a self-reinforcing dynamics where increased norm violation decreases norm acceptance, leading to even more norm violation and so on. As demonstrated by the equilibrium analyses (see center panel of Fig. 1), these model dynamics move the population increasingly close to a state where all agents have a minimal norm acceptance and a very high probability of norm violation. Already at $t=150$, average norm acceptance had adopted a very small value of .004 ($sd=.005$) and the smallest probability of norm violation was .99994. This trend continued until the probability of norm violation of all agents approached a value of one. This extreme outcome could only be reached because norm acceptance was flexible. Otherwise, norm acceptance would have remained being uniformly distributed leading to much lower probabilities of norm violation (Eq. (1)).

An important outcome of the simulation experiment was that the dynamics that Fig. 5 illustrates unfold only when the social network is sufficiently clustered, because populations with a low degree of clustering reach a maximal or a minimal rate of norm violation already before the disclosure at $t=100$. Fig. 6 plots the rate of norm violation that we observed just before the true rate was disclosed ($t=99$) and how this was affected by the degree of network clustering (parameter R). It shows that the rate of norm violation adopted moderate levels in most simulation runs with

Fig. 5. Ideal-typical dynamics when norm acceptance is flexible. ($\delta = 10$; $R = 0$; $\alpha = 1$; $\gamma = 0.2$, repeated disclosure).

Fig. 6. The effect of network clustering on the rate of norm violation observed before the disclosure (norm acceptance is flexible, $\alpha = 1$).

clustered networks ($R < 20$). The smaller clustering was, however, the more often population consisted either of only norm violating or only norm complying agents already before the disclosure of the true rate.

The finding that low network clustering facilitates coordination is in line with insights from the literature on social-influence dynamics, where it has been demonstrated that including network ties between clusters facilitates convergence toward a global consensus (Flache and Macy, 2011). In highly clustered networks, agents form transitive communities where connected agents are influenced by the same set of others. Thus, influence from the outside is relatively small and local biases toward a low or a high norm acceptance can persist. What is more, these differences can intensify in our model, because high (low) norm acceptance leads to low (high) norm violation, which, in turn, increases (decreases) norm acceptance. This explains why in clustered networks internally homogeneous but mutually maximally different subsets of agents can form. In networks with a large number of rewired links,

in contrast, agents are connected to others who are influenced by a different set of agents. This diverse influence makes it very likely, that all agents develop similar degrees of norm acceptance. In these networks, we did not observe the bi-polarized distributions of norm acceptance that Fig. 5 shows. Instead, the distribution of norm acceptance shifts from a uniform distribution at the outset to a normal distribution. Next, agents collectively shift either to maximal or minimal degrees of norm acceptance.

The extremely fast formation of consensus in networks with low clustering took us by surprise and interfered with the design of the experiment, where the disclosure starts at $t = 100$. We therefore tested whether the rate of norm violation can be disclosed earlier in the simulation runs. However, it turned out that a maximal or a minimal rate of norm violation was reached within very few periods even when social influence was assumed to be even weaker than in the experiment ($\gamma < .2$). What is more, runs with a disclosure already at the very beginning of the simulations were difficult to interpret, as the observed dynamics resulted from both the disclosure and the general trend toward either maximal or minimal norm violation that social influence generates, making it difficult to quantify the disclosure effect.

In sum, we interpret the observed effects of network clustering in the following way. The assumption of peer influence implies that populations will always develop a consensus on norm acceptance, which formal models of social influence have demonstrated (Abelson, 1964; DeGroot, 1974). In networks with a small degree of clustering, this convergence process can be so fast that disclosing the rate of norm violation would not alter the dynamics. In networks with a high degree of clustering, in contrast, the convergence process that social influence generates can take much longer and disclosing the true rate of norm violation can drastically affect behavioral dynamics as Fig. 5 illustrates. In sum, an important and so far overlooked insight from our modeling exercise is that when norm acceptance is open to peer influence then disclosure effects can only obtain in clustered networks.

Figs. 7–9 show the results of the simulation experiment for the conditions where norm acceptance was flexible ($R = 5$; $\gamma = .2$), using the same methods as Figs. 2–4. The figures show that norm violation changed significantly also in the runs without any disclosure. These

Fig. 7. Development of the distribution of rate of norm violation in the 500 replications per disclosure scenario condition with flexible norm acceptance.

changes reflect that connected subgroups of agents with the same norm acceptance (see box for $t=99$ in Fig. 5) are not stable and may disappear when the neighboring subgroups grow. As it is more likely that small subgroups disappear than big ones, the average rate of norm violation tends to either increase or decrease over time.

Fig. 7 shows that the effects of disclosing the true rate of norm violation were much stronger than in the simulations with fixed norm acceptance. In fact, all simulation runs where agents had a memory or where the true rate was disclosed repeatedly quickly reached a state where either all agents violated the norm or all agents abided by the norm. Particularly interesting is the gradual change in norm violation in the runs with single disclosure and decreasing memory. This gradual change demonstrates that the disclosure of the true rate did not only have short-term effects on agents' probability to violate the norm. In addition, the disclosure sparked cascades of behavioral changes that clearly exceeded the change in norm violation measured directly after the disclosure

($t=100$). This supports our prediction that cascades can alter norm violation over and above the short-term disclosure effects that the exiting literature focused on.

Comparing the panels in Fig. 7 for the conditions where agents had a decreasing memory with those where agents had a long memory provides further evidence for the impact of cascades. The figure shows strong disclosure effects in the runs with long memories, but most of these runs did not reach maximal or minimal rates of norm violation while the agents remembered the disclosed rate. This happened because agents developed a norm acceptance equal to one minus the rate of norm violation at $t=99$. As the disclosed rate of norm violation was not updated, also norm acceptance remained constant. This, in turn, implied that also the rate of norm violation did not adopt extreme values. In contrast, with a decreasing memory, the disclosure effects were much stronger because a self-reinforcing dynamic developed. After the disclosure, agents adjusted their norm acceptance and their behavior. Next they experienced similar changes in the behavior of their neighbors and again adjusted their norm acceptance and behavior. This process ended when either all agents violated the norm or all agents followed the norm. This self-reinforcing dynamics could not unfold when agents had a long memory. However Fig. 7 shows that at the moment when agents in the scenario with a long memory forgot the disclosed rate ($t=200$), the same self-reinforcing dynamics unfolded.

Similar to what we observed in the simulations with fixed norm acceptance, the disclosure of the true rate of norm violation did not always lead to an increase in norm violation. In fact, norm violation increased as often as it decreased (see average lines in Fig. 7). Furthermore, Fig. 9 shows that norm violation increased when the true rate of norm violation was above 0.5 and decreased otherwise, which answers Research Question 1.

Did dynamics differ when the rate of norm violation was disclosed once or repeatedly? Figs. 7 and 8 show that in the simulations with the single disclosure and no memory, norm violation after the disclosure differed only slightly from norm violation in the simulation without any disclosure. This small difference, however, is statistically significant (see Fig. 8). For the other disclosure scenarios, however, Figs. 7–9 show very strong disclosure effects and that the effects were much more pronounced than in the simulations with fixed norm acceptance.

3.3. Sensitivity tests

We tested the sensitivity of our findings to changes in the amount of noise in agents' behavior (δ), and the skewness of the initial distribution of norm acceptance (α) and found that the results are robust to changes in these parameters. Average changes in norm violation were stronger when there was more noise (low δ) in the simulations without any disclosure and under those conditions where the disclosure had only limited effects (e.g. single disclosure with no memory). Thus, under these conditions changes resulted from random fluctuations. In those conditions where we observed significant disclosure effects, however, noise had the opposite effect, leading to weaker changes in norm violation, which shows that the reported findings are more pronounced when agents' behavior follows from our model assumptions rather than randomness.

Results were also robust to changing the initial distribution of norm acceptance (α). Of course there was less norm violation at the moment of the disclosure when the initial distribution of norm acceptance was negatively skewed. As a consequence, norm violation most often decreased after the disclosure. Furthermore, changes in norm violation were smaller when initial norm acceptance was skewed, because norm violation adopted very small values from the beginning on and could decrease only slightly after the disclosure.

Fig. 8. Results of the simulation experiment with flexible norm acceptance (based on all runs of the experiment with $\alpha = 1$), error bars show 95 percent confidence intervals.

4. Summary

Heinrich Popitz formulated the intriguing hypothesis that disclosing a population's rate of norm violation will lead to increased norm violation and norm decay. He concluded that there is a *preventive effect of ignorance* (Popitz, 1968, 2006) as increasing norm violation can be prevented if the population remains uninformed about the true degree of norm violation. We studied a threshold model to derive new hypotheses about the structural conditions under which disclosing the true rate of norm violation increases or decreases collective norm violation. Contrary to the existing literature on disclosure effects, we took into account that individuals are not isolated but influence each other's behavior. Influence can lead to cascades that profoundly change collective behavior over and above the disclosure effect on isolated individuals that earlier studies reported.

Our analyses show that the model's dynamics and its predictions concerning the consequences of disclosing the true rate of norm violation depend very much on whether thresholds are fixed or open to peer influence. Fixed thresholds correspond to settings where norm acceptance, the degree to which individuals intrinsically value norm compliance, is not changing over time. It turned out that disclosure effects are very limited in this case. In fact, we found substantial changes in collective norm violation only when the rate of norm violation was measured and disclosed repeatedly. However, even these effects obtained only in the short run and leveled off after the rate of norm violation was not disclosed any longer. In the short run, however, disclosing the true rate, individuals' reactions, and the disclosure of the updated rate generated a self-reinforcing dynamics that results in strong changes in norm violation. This leads to the first new hypothesis, which provides an answer to Research Questions 3 and 4.

Hypothesis 1. In settings with inflexible norm acceptance, disclosing the true rate of norm violation has very limited effects, unless norm violation is measured and disclosed repeatedly.

When thresholds are flexible, the model can generate very strong disclosure effects. Flexible thresholds correspond to settings where norm acceptance is intensified when individuals observe

low norm violation and norm acceptance decreases when actors believe that norm violation is high. In simulations with flexible thresholds, we found strong disclosure effects only when the social network was highly clustered. In networks with low clustering, peer influence leads to a rapid emergence of consensus on norm acceptance and uniformity in norm violation, a model dynamics that mirrors classical findings from models of social influence (Abelson, 1964). Even though peer influence was assumed to be rather weak, this convergence process was so rapid that disclosing the true rate of norm violation hardly affected the dynamics.

Hypothesis 2. When norm acceptance is open to peer influence, disclosure effects can only obtain in sufficiently clustered networks.

However, even when norm acceptance is open to peer influence and when the network is clustered, disclosure effects are very limited when the true rate is disclosed only once and agents consider it only in the period right after the disclosure. This is an important finding, as it shows that the collective change in norm violation is limited even though many individuals might change their behavior. The problem is that there are always agents who overestimate the rate of norm violation and agents who underestimate it. The behavioral changes of the two sets of actors often balance out one another and the overall disclosure effect remains small.

Our analyses demonstrate that the model with flexible thresholds predicts very strong disclosure effects when agents consider the disclosed information over a longer period or when the rate of norm violation is measured and disclosed repeatedly. In fact, it turned out that all simulations under these conditions ended in a state where either all agents violated the norm or norm violation was zero as a result of the disclosure. The mechanism underlying this strong effect is a self-reinforcing dynamic. If, for instance, a cluster of norm abiding agents learns that the global rate of norm violation is higher than in their network neighborhood, then these agents are going to decrease their norm acceptance. The immediate result is that also within the cluster norm violation will increase, which will be observed by the agents. These observations, in turn, lead to a further decrease in norm acceptance and so on. Obviously, these dynamics cannot unfold when agents always base their

Fig. 9. Direction of the change in the rate of norm violation as a function of the rate of norm violation just before disclosure when norm acceptance is flexible (based on all runs of the experiment with $\alpha = 1$ and $\delta = 10$).

estimate of the rate of norm violation on the disclosed value (scenario with single disclosure and long memory), leading to the somewhat counter-intuitive prediction that the disclosure effect is stronger when individuals have a limited memory. This outcome is interesting for the design of possible intervention strategies, as it suggests that repeatedly informing the population about the norm violation at a given point in time ($t=99$) weakens the disclosure effect.

Hypothesis 3. When norm acceptance is open to peer influence, disclosure effects are weaker when individuals have long

memories in the sense that they base their estimate of the rate of norm violation always on rate of norm violation that was disclosed.

Research Question 1 asked how the magnitude of the disclosed rate affects whether the rate of norm violation increases or decreases. In turned out that the disclosure tends to reinforce the dominant behavior in the population. In other words, when norm violation exceeds 50 percent, then the disclosure further intensifies norm violation. When fewer than 50 percent of the agents violate the norm, the disclosing this information leads to even less norm violation.

Hypothesis 4. If there is a disclosure effect, then it tends to reinforce the dominant behavior in the population.

5. Discussion

Our results support Popitz's conclusion that there might be a preventive effect of ignorance. However, our four new hypotheses show that this effect will obtain only under specific conditions. Obviously, the next step is to put the new hypotheses to the test.

Experiments in the laboratory and the field are a very promising approach to empirically test the four hypotheses. A great advantage of the laboratory is that one cannot only test the hypotheses but also important underlying model assumptions. For instance, [Rauhut \(2013\)](#) elicited participants' estimates of the rate for norm violation in his experiment, making it possible to test whether or not these estimates translate into behavior as assumed by the model. As our analyses demonstrate, model predictions depend critically on whether norm acceptance is flexible or not. Measuring norm acceptance before and after the disclosure would allow to test our implementation of this crucial model ingredient.

Furthermore, field experiments are promising if they, contrary to many existing field experiments on norm violation (e.g. [Keizer et al., 2008](#)) go beyond the measurement of direct responses by isolated individuals (e.g. [Centola, 2010](#)). In order to test effects of cascades, individuals need to be given the chance to observe behavioral changes in others and adjust their own behavior multiple times, which is challenging but not impossible. For instance, a typical social norm in many organizations is to save on paper. To repeatedly measure the degree of norm violation in different divisions of an organization, one can extract from the print logs of each printer the share of double-sided print jobs. Popitz's prediction would be that the share of double-sided print jobs decreases when employees are informed about the global rate of norm violation. In this context, it is even possible to manipulate network structure ([Hypothesis 2](#)) by informing participants prior to disclosing the rate of norm violation about the average rate of norm violation at the printers close to their own offices (high clustering) or a random set of printers (low clustering).

Future work should also elaborate the theoretical model. For instance, we assumed that individuals correctly observe local and global rates of norm violation. This may be an acceptable assumption if exact numbers of norm violation are disclosed, but often citizens are only informed that norm violation occurred more frequently than expected. For instance, in 2009/2010 the media disclosed only cases of child abuse by priests, rather than the exact rate of child abuse. Future work should explore whether predictions change, when agents are provided with more ambiguous information.

Future work should also study more network effects. To be able to study effects of network clustering, we used a rewiring algorithm that keeps constant other important network characteristics, such as the degree distribution. As an agent's degree can influence the accuracy of her estimate of collective norm violation, it would be interesting to explore dynamics in networks with different degree distributions.

Furthermore, future research should directly model the mechanisms underlying individuals' reactions to the disclosure. For instance, it is certainly plausible that individuals increase norm violation after the disclosure because they believe that the high degree of norm violation was rooted in low norm sanctioning. It might be interesting to include sanctioning directly in the model, because sanctioning expectations might be more or less strong. In the extreme case, expected sanctioning will not change, for example, when a high number of unexpectedly frequent tax frauds are disclosed together with the information that the tax authorities hired more enforcement personnel. In other cases the belief might originate that negative sanctions are very unlikely. Future modeling work should study how model predictions depend on heterogeneous sanctioning beliefs and behavior.

All models proceed from simplifying assumption. Methodologically, this is a key tool of theory development: important variables are singled out and other factors are assumed to not influence the given dependent variables. These other factors are scope conditions for the validity of the model. Any theory makes such assumptions but too often those assumptions are not explicitly formulated. Since formal models must lay open what is assumed it is easier to see which scope conditions are assumed. Thus, the existence of such conditions is not so much a weakness of the model but an incentive for further research.

Acknowledgements

Mäs acknowledges support by the European Commission through the ERC Advanced Investigator Grant 'Momentum' (Grant 324247). We thank the two anonymous reviewers, Andreas Flache, Jacob Dijkstra, Andreas Diekmann, Joel Berger, Thomas Grund, Olivia Woolley-Meza, Bary Pradelski, and Matthias Andreas Näf, as well as the members of the Norms and Networks cluster at the University of Groningen for helpful comments.

References

- Abelson, R.P., 1964. Mathematical models of the distribution of attitudes under controversy. In: Frederiksen, N., Gulliksen, H. (Eds.), *Contributions to Mathematical Psychology*. Rinehart Winston, New York, pp. 142–160.
- Allen, R.E., 2008. *The Republic of Plato*. Yale University Press, New Haven.
- Barash, V., Cameron, C., Macy, M., 2012. Critical phenomena in complex contagions. *Soc. Netw.* 34 (4), 451–461.
- Cabinet-Office, 2012. *Applying Behavioural Insights to Reduce Fraud, Error and Debt*. Cabinet Office. Behavioral Insights Team, London.
- Centola, D., 2010. The spread of behavior in an online social network experiment. *Science* (80-) 329 (5996), 1194–1197.
- Centola, D., Macy, M., 2007. Complex contagions and the weakness of long ties. *Am. J. Sociol.* 113 (3), 702–734.
- DeGroot, M.H., 1974. Reaching a consensus. *J. Am. Stat. Assoc.* 69, 118–121.
- Diekmann, A., Przepiorka, W., Rauhut, H., 2015. Lifting the veil of ignorance: an experiment on the contagiousness of norm violations. *Ration. Soc.* 27 (3), 309–333, <http://rss.sagepub.com/content/27/3/309.short> (retrieved 11.03.16).
- Epstein, J.M., 2007. *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton University Press, Princeton.
- Festinger, L., Schachter, S., Back, K., 1950. *Social Pressures in Informal Groups*. Stanford University Press, Stanford, CA.
- Flache, A., Macy, M.W., 2011. Small worlds and cultural polarization. *J. Math. Sociol.* 35 (1), 146–176.
- Franck, G., 1999. Scientific communication – a vanity fair? On JSTOR. *Science* (80-) 286 (5437), 53–55, http://www.jstor.org.proxy-ub.rug.nl/stable/2898883?seq=2#page_scan_tab_contents (retrieved 11.03.16).
- Friedkin, N.E., Johnsen, E.C., 2011. *Social Influence Network Theory*. Cambridge University Press, New York.
- Funk, S., Salathe, M., Jansen, V.A.A., 2010. Modelling the influence of human behaviour on the spread of infectious diseases: a review. *J. R. Soc. Interface* 7 (50), 1247–1256.
- Granovetter, M., 1978. Threshold models of collective behavior. *Am. J. Sociol.* 83 (6), 1420–1443.
- Groeber, P., Rauhut, H., 2009. Does ignorance promote norm compliance? *Comput. Math. Organ. Theory* 16 (1), 1–28.
- Hegselmann, R., Krause, U., 2002. Opinion dynamics and bounded confidence models, analysis, and simulation. *J. Artif. Soc. Soc. Simul.* 5 (3).
- Hoffmann, M., Lauer, T., Rockenbach, B., 2013. The royal lie. *J. Econ. Behav. Organ.* 93, 305–313.
- Keizer, K., Lindenberg, S., Steg, L., 2008. The spreading of disorder. *Science* (80-) 322 (5908), 1681–1685.
- Keizer, K., Lindenberg, S., Steg, L., 2013. The importance of demonstratively restoring order. *PLOS ONE* 8 (6).
- Keuschnigg, M., Wolbring, T., 2015. Disorder, social capital, and norm violation: three field experiments on the broken windows thesis. *Ration. Soc.* 27 (1), 96–126.
- Krassa, M.A., 1988. Social groups, selective perception, and behavioral contagion in public opinion. *Soc. Netw.* 10 (2), 109–136.
- Kuran, T., 1989. Sparks and prairie fires – a theory of unanticipated political revolution. *Public Choice* 61 (1), 41–74.
- Lopez-Pintado, D., Watts, D.J., 2008. Social influence, binary decisions and collective dynamics. *Ration. Soc.* 20 (4), 399–443.
- Macy, M.W., 1991. Chains of cooperation – threshold effects in collective action. *Am. Sociol. Rev.* 56 (6), 730–747.
- Macy, M.W., Willer, R., 2002. From factors to actors: computational sociology and agent-based modeling. *Annu. Rev. Sociol.* 28 (1), 143–166.
- Macy, M., Flache, A., 2009. Social dynamics from the bottom up. Agent-based models of social interaction. In: Hedström, P., Bearman, P. (Eds.), *The Oxford Handbook of Analytical Sociology*. Oxford University Press, Oxford, pp. 245–268.
- Maslov, S., Sneppen, K., 2002. Specificity and stability in topology of protein networks. *Science* (80-) 296 (5569), 910–913.
- McFadden, Daniel, 1973. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), *Frontiers in Econometrics*. Academic, New York, pp. 105–142.
- Meade, N., Islam, T., 2006. Modelling and forecasting the diffusion of innovation – a 25-year review. *Int. J. Forecast.* 22 (3), 519–545.
- Oliver, P., Marwell, G., Teixeira, R., 1985. A theory of the critical mass. 1. Interdependence, group heterogeneity, and the production of collective action. *Am. J. Sociol.* 91 (3), 522–556.
- Opp, K.-D., 2011. The beneficial and unintended consequences of false beliefs about norm violation. When is there a 'preventive effect of ignorance'? In: Mica, A., Peisert, A., Winczorek, J. (Eds.), *Sociology and the Unintended*. Robert Merton Revisited. Peter Lang, Frankfurt am Main, pp. 257–283.
- Popitz, H., 1968. Über Die Präventivwirkung Des Nichtwissens. *Dunkelziffer, Norm Und Strafe*. Mohr Siebeck, Tübingen.
- Popitz, H., 2006. In: Eßbach, W., Pohlmann, F. (Eds.), *Soziale Normen*. Suhrkamp, Frankfurt am Main.
- Rauhut, H., 2013. Beliefs about lying and spreading of dishonesty: undetected lies and their constructive and destructive social dynamics in dice experiments. *PLOS ONE* 8 (11).
- Rogers, E.M., 1995. *Diffusion of Innovations*. Free Press, New York.
- Schelling, T.C., 1978. *Micromotives and Macrobehavior*. WW Norton and Company, New York.
- Schultz, P.W., Nolan, J.M., Cialdini, R.B., Goldstein, N.J., Griskevicius, V., 2007. The constructive, destructive and reconstructive power of social norms. *Psychol. Sci.* 18 (5), 429–434.
- Thaler, R.H., Sunstein, C.R., 2009. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Penguin Books, New York.
- Valente, T.W., 1996. Social network thresholds in the diffusion of innovations. *Soc. Netw.* 18 (1), 69–89.
- Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. *Nature* 393 (6684), 440–442.
- Wood, W., 2000. Attitude change: persuasion and social influence. *Annu. Rev. Psychol.* 51, 539–570.