HARVARD | BUSINESS | SCHOOL

Testing Coleman's Social-Norm Enforcement Mechanism: Evidence from Wikipedia

Mikołaj Jan Piskorski
Andreea Gorbatai

Working Paper

11-055

March 26, 2013

TESTING COLEMAN'S SOCIAL-NORM ENFORCEMENT MECHANISM: EVIDENCE FROM WIKIPEDIA[*]

MIKOŁAJ JAN PISKORSKI
Harvard University
Morgan Hall 243
Harvard Business School
Boston, MA 02163

Tel. (617) 495-6099

mpiskorski@hbs.edu

ANDREEA GORBATAI
Harvard University
Morgan Hall T37
Harvard Business School
Boston, MA 02163

Tel. (617) 823-9388

agorbatai@hbs.edu

[*] Both authors contributed equally to this article; their names are listed in reverse alphabetical order. We are grateful to Isabel Fernandez-Mateo and Peter Marsden and to participants in the Academy of Management 2009, INSNA 2009, and WOM 2008–2009 seminar for their comments on this paper. John Sheridan helped us assemble the dataset. The Division of Research at Harvard Business School provided financial support. All errors are ours.

TESTING COLEMAN'S SOCIAL-NORM ENFORCEMENT MECHANISM: EVIDENCE FROM WIKIPEDIA

ABSTRACT

Since Durkheim, sociologists have believed that dense network structures lead to fewer norm violations. Coleman (1990) proposed one mechanism generating this relationship and argued that dense networks provide an opportunity structure to reward those who punish norm violators, leading to more frequent punishment and in turn fewer norm violations. Despite ubiquitous scholarly references to Coleman's theory, little empirical work has directly tested it in large-scale natural settings with longitudinal data. We undertake such a test using records of norm violations during the editing process on Wikipedia, the largest user-generated on-line encyclopedia. These data allow us to track all three elements required to test Coleman's mechanism: norm violations, punishments for such violations and rewards for those who punish violations. The results are broadly consistent with Coleman's mechanism.

Introduction

Sociologists have long invoked norms to explain social order (Durkheim 1893; Parsons 1953) and to account for various aspects of social behavior (Weber 1976). Norms embody a group's social consensus about appropriate behaviors. Some norms prohibit behaviors deemed unacceptable and specify punishments for flouting these proscriptions (Homans 1950:123). Others prescribe behaviors and reward those who undertake them (Blake and Davis 1964).

Among the various types of norms, sociologists have taken a particular interest in social norms. These norms require that parties personally unaffected by norm violation either punish offenders (Coleman 1990), or reward those who conform (Goode 1978). These characteristics of social norms raise a fundamental question: why one actor would punish or reward another for actions affecting others (Horne 2004). Since these rewards and punishments are costly for those who mete them out, but largely benefit others, potential enforcers are likely to have insufficient incentives to enforce norms (Olson 1971). In the absence of such enforcement, social norms will not be observed (Coleman 1990; Oliver 1980).

Scholars from many disciplines have examined factors that lead people to enforce such social norms (e.g. Axelrod 1986; Bendor and Swistak 2001; Ellickson 2001; Fehr and Gächter 2002). Among these explanations, sociologists have particularly invoked the role of network density (Burt 1982; Durkheim 1951; Lin 2001; Simmel 1902:170). Coleman (1990) formalized the argument, theorizing that high-density networks provide an opportunity structure within which third parties can compensate norm enforcers for the expense of chastising norm violators. Such payments encourage actors to punish those who violate norms, which in turn reduces the incidence of norm violation.

Judging by the number of citations, Coleman's argument is now taken for granted in sociology (Horne 2001; Morgan and Sørensen 1999; Sampson, Raudenbush, and Earls 1997). There

is also ample evidence of a negative correlation between network density and norm violations across numerous settings. Researchers have argued, for example, that norms against malfeasance among diamond traders and among geographically dispersed medieval Maghribi traders were sustained by high-density networks (Coleman 1990; Greif 1989). Similarly, in rotating-credit and informal help associations, the ability to sustain the norm of contributing to others' welfare has been shown to be associated with high density among associations' members (Barker 1993; Biggart 2001; Coleman 1989:s102-03; Uehara 1990).

However, there is reason for skepticism that such correlational evidence can be used to support a causal link between network density and infrequent norm violations. First, some of the studies cited above examine a single social system and make inferences by pointing to the co-presence of network density and absence of norm violations without showing the counterfactual. Other studies that have undertaken comparative design were largely cross-sectional, making it difficult to establish causality. Furthermore, existing work provides little evidence to support Coleman's mechanism. This is problematic because simpler explanations can generate the same empirical predictions (Elster 2003). Consider, for example, a high-mutual-dependence environment in which actors exchange resources they value highly (Molm 1997). It is easy to see that actors in such environments will violate norms infrequently, and that they will also establish dense relationships with one another (Horne 2001). In this case the relationship between density and norm violations is not causal but arises out of high mutual dependence (Flache and Macy 1996).

To provide evidence for Coleman's mechanism, it is necessary to follow his three-step reasoning process, and to provide support for each step using longitudinal data. Specifically, it is first necessary to confirm that norm violations decline as network density increases. Second, a researcher needs to furnish evidence that higher network density leads to more actions eliciting norm

compliance (such as punishing norm violations), which then will lead to lower norm violations. The third step is to show that higher network density leads to more acts of compensating those who elicit norm compliance, which then leads to more acts of eliciting norm compliance, which then leads to fewer norm violations. Without supporting all three assertions, it is hard to assert that Coleman's argument has been tested properly. In this paper, we perform all three tests.

We undertake them in the context of the community of editors of Wikipedia, the largest user-generated on-line encyclopedia (Anthony, Smith, and Williamson 2009). This setting allows us to study norm violations in a naturalistic setting but at the same time to clearly observe (1) who violated a norm and who suffered from the violation, (2) who, if anyone, stepped in to punish, and (3) whether those who punished norm violators received rewards from the community for doing so. Also, because we observe actors over time as they experience transitions from a dense network to a sparse one (or vice-versa), we can provide results that are subject to fewer alternative interpretations. Finally, the network relationships we study are fairly weak, therefore providing very conservative tests of Coleman's theory.

The remainder of the paper is structured as follows. The next section examines the existing literature on norms, with particular emphasis on Coleman's mechanism, to derive our key hypotheses. We next describe our setting and data, and then our results. The final section discusses the limitations of our study and its conclusions.

Theory

Step 1: Violating norms

A norm is a set of rules specifying appropriate behaviors and backed by social rewards or sanctions (Blake and Davis 1964). Norms can be characterized on three dimensions. First, norms differ in their

valence. Prescriptive norms encourage given actions, such as clapping at the end of a performance; proscriptive norms discourage specific actions, such as carrying a loaded gun. Second, norms differ in the types of behaviors they seek to regulate. Certain norms, often called *conventional*, seek to make everyone choose a single coordinated form of action that benefits all. Driving on the same side of the road is a conventional norm. Other types of norms resolve conflicts of interest between individuals and others. Often called *essential*, these norms mandate behavior that is beneficial to others but costly to the individual. Essential norms also prohibit behavior harmful to others but gratifying to the individual (Hechter 1987; Hechter and Kanazawa 1993). The norm not to pollute urban streets, for example, is beneficial to everyone but requires individuals to carry their trash rather than disposing of it on the spot.

It is easier to explain theoretically why actors comply with conventional norms than with essential norms. Because conventional norms are in everyone's interest, and no individual benefits arise from violating them, self-interested actors will comply with conventional norms. It is harder to understand why such actors comply with essential norms, since they bear the individual costs of compliance but appropriate only part of the benefit. This scenario leads to a (first order) free-rider problem whereby every actor prefers not to comply with an essential norm but wants everyone else to do so. If all actors reason this way, no one will follow the norm. Thus, theoretical formulations of essential norms need to account for why self-interested actors comply with such norms.

Finally, norms also differ with regard to whether those who are expected to comply with them benefit from such behavior. Norms that benefit those who adhere to them are often called *conjoint*. A norm restricting use of a single telephone in a dormitory to ten minutes would fall into this category. At the other extreme are norms that do not benefit those who adhere to them, instead benefitting another group; such norms are usually called *disjoint*. An example is children who are

expected to behave appropriately for the benefit of their adult caretakers. Most norms fall somewhere between the two ends of this spectrum, benefitting both those who comply with the norm and others who are not subject to it.

The distinction between conjoint and disjoint norms has implications for the free-rider problem associated with essential norms. In the case of conjoint norms, those who incur the cost of observing the norm are also its beneficiaries. Thus the free-rider problem is present but contained to a certain degree by the fact that individuals derive some of the benefits of their own normative behaviors. In the case of disjoint norms, those who incur the cost of following a norm are not its beneficiaries. Such absence of direct benefits accentuates the free-rider problem. This implies that it will be even more important for the beneficiary group to elicit norm compliance from the target group.

Step 2: Eliciting norm compliance

Given the difficulty of eliciting norm compliance, it is important to understand when and how it occurs (Bendor and Swistak 2001; Ellickson 1991; Homans 1950:123; Yamagishi and Cook 1993). In general, it is costly to elicit norm compliance. Resources used as rewards or punishments cannot be used for another purpose, but those who use their resources to elicit compliance enjoy only a fraction of its benefits. For most actors, the expected benefit will be too small relative to the cost; thus each actor will wait for others to elicit compliance. But if all potential actors behave in this way, no one will seek to elicit norm compliance. Coleman (1990) called this phenomenon "the second-order free-rider problem" to distinguish it from the first-order free-rider problem of compliance with norms described in Step 1.

The severity of the second-order free-rider problem depends on how actors seek to elicit norm compliance. Sometimes compliance is elicited with rewards. Goode (1978) argued, for example, that status can be used as a payment for complying with norms, particularly prescriptive norms. In other cases failure to comply with a norm elicits punishment, such as public chastisement. The distinction is important, because rewarding others rarely elicits negative reactions, whereas punishing others can easily prompt retaliation by those punished. As a consequence, actors are less likely to punish than to reward others (Molm 1997), and so the second-order free-rider problem is accentuated when punishment is used to elicit norm compliance (Horne 2007). It is thus particularly important to compensate those who punish others for failing to observe norms, a topic we will return to in Step 3 later.

Second, eliciting norm compliance can take the form of group or individual effort. When a group seeks to elicit norm compliance, each member can provide a small part of the reward or the punishment at a reasonably modest cost. Since the cost of eliciting norm compliance by a group is small, the second-order free-rider problem is attenuated (but not eliminated). In contrast, when a single individual is entirely responsible for eliciting norm compliance, he bears the entire cost and the second-order free-rider problem is accentuated. Thus, compensating those who individually elicit norm compliance is particularly important.

Finally, norm compliance can be enforced either by those affected or by unaffected third parties. In most Western societies, for example, parents alone are expected to punish their misbehaving young children. Since those directly affected by the norm have a greater incentive to elicit compliance, the second-order free-rider problem is attenuated. For other norms, however, parties unaffected by a norm transgression are expected to step in and punish the offender. Norms backed by enforcement of this kind are often called *social norms*. For example, publicly

disapproving of someone who fails to give up a seat on a bus for an elderly or a handicapped person is a social norm; unaffected third parties are expected to chastise someone who refuses to do so. Because such third parties bear the entire cost of eliciting norm compliance and appropriate none of the benefits, the second-order free-riding problem is quite strong. In such situations, compensating third parties for eliciting norm compliance is particularly important.

Step 3: Compensating those who elicit norm compliance[1]

Despite the difficulties of eliciting norm compliance, it is possible to compensate those who engage in such acts. As before, compensating those who elicit norm compliance is subject to another free-rider problem, often called "the third-order free-rider problem" (Elster 2003; Horne 2001). The problem arises because such compensation is costly. Thus, each actor waits for others to provide compensation so as to appropriate the benefits without incurring the costs. This problem is most pronounced when compensation needs to be provided for punishing norm violators. As suggested above, eliciting norm compliance via punishment is more expensive than doing so via rewards, calling for a higher level of compensation.

A number of theories have sought to solve this third-order free-rider problem of compensation for punishing norm violations. Some theories invoked the intrinsic satisfaction derived from compensating others for eliciting norm compliance (Knutson 2004). Others suggested that bestowing rewards or punishments on those who compensate could solve the third-order free-rider problem. But this approach generates a fourth-order free-riding problem, leading to an infinite-regress problem (Elster 1989). To avoid this problem, most theories focused on solving the third-order free-rider problem directly. Specifically, a broad set of theories suggested that the third-order

[1] We use the word *elicit* to designate the act of inducing others to observe a norm, and *compensate* to designate the act of inducing others to elicit norm compliance. Thus, compensating logically precedes eliciting. Both actions can take the forms of giving rewards or administering punishments.

free-rider problem can be overcome when punishments for norm violations are compensated with rewards. This line of reasoning has gained substantial acceptance in the extensive evolutionary literature on norms, which shows that the rule of rewarding those who punish deviance wins in competition with other behavioral rules (Bendor and Swistak 2001; Opp 1982; Schotter 1982; Sugden 1986). In the same vein, most experimental results show that actors are most likely to observe norms when those who sanction norm violators are rewarded (Horne 2001). Finally, Coleman (1990) argued that compensation through rewards is less likely to suffer from the third-order free-riding problem, because rewards are cheaper to furnish than punishments.[2] Once this third-order free-riding problem is solved this way, Coleman argued that it is possible to solve both the second- and first-order problems and thus ensure that norms are observed.

Coleman's solution to the free-rider problems

To understand Coleman's argument consider a numerical example with three actors, A, B and C. Assume that actor A considers whether to disobey a norm, which would bring personal benefits of $30 to A, but would also impose a cost of $30 on actor B and C each. This creates the first-order free rider problem, and will lead actor A to violate the norm unless he thinks he might be punished. Actor B or C could punish actor A for violating the norm, but assume that each would have to incur a cost of $35 to do so. If that's the case, neither actor B nor actor C will punish actor A. This leads to the second-order free rider problem. In principle, one of the affected actors, say C, could reward another, say B, for punishing actor A. However, actor B would have to receive a reward of, say, $40 to compensate him for the $35 cost to punish actor A. If actor C needs to incur substantial cost to provide this reward, say, also $40, he is he is unlikely to provide such a reward.

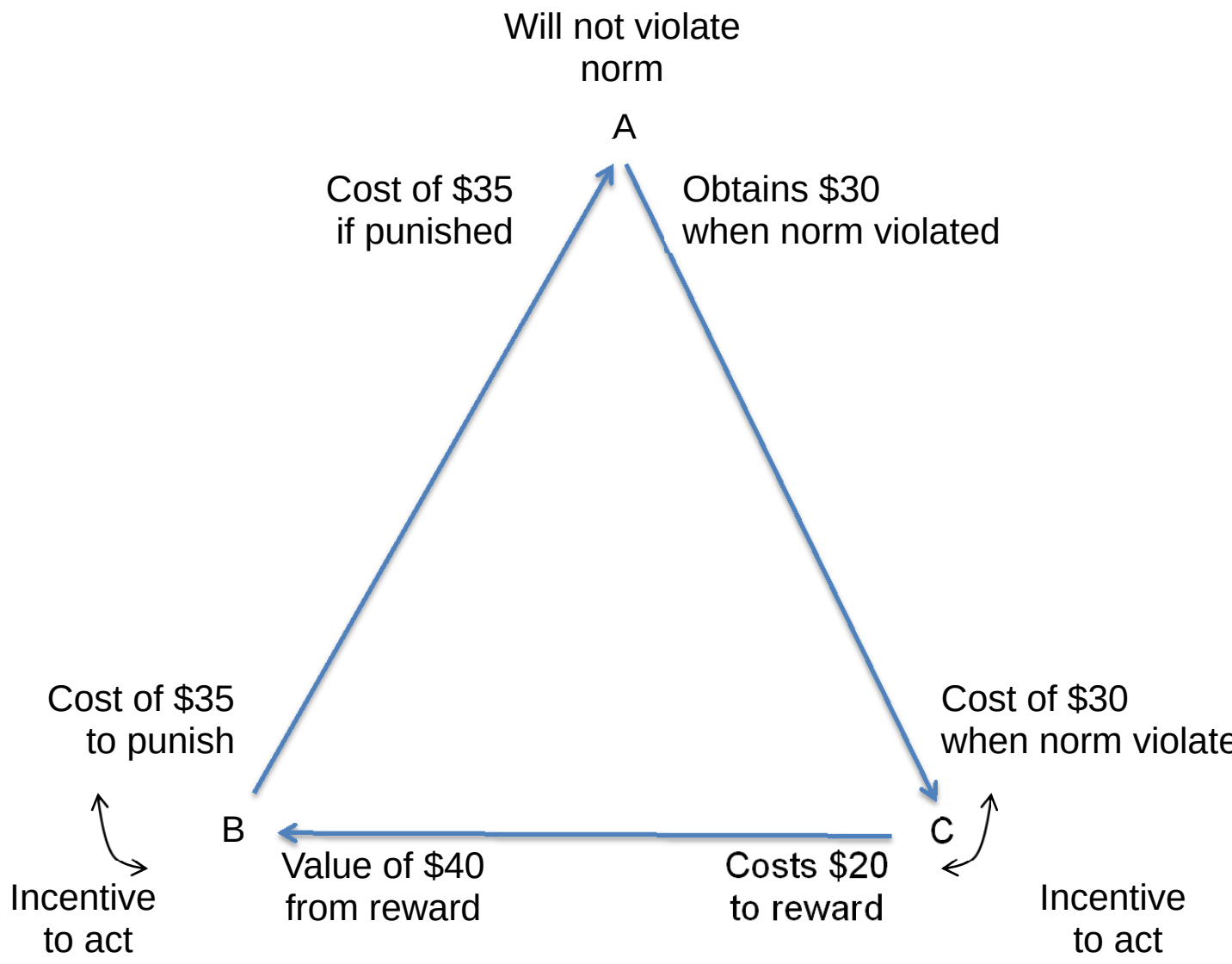
[2] Coleman (1990:283) captured this argument stating: "Where sanctions are applied in support of a proscriptive norm and are consequently negative sanctions, the . . . problem of providing positive sanctions for the sanctioner is more easily overcome, because positive sanctions incur lower costs than do negative ones."

He would rather suffer the $30 cost associated with norm violation than provide the $40 reward to B. The same logic applies to B rewarding C, leading to the third-order free rider problem.

Coleman argued, however, that there is a class of rewards that are very valuable to actor B, but cheap for actor C to furnish, as illustrated in **Diagram 1** below.[3] Suppose that such rewards only cost $20 to actor C, but give $40 value to actor B. In this scenario, actor C will be willing to incur the cost of $20 to give such a reward, because he can avoid the cost of $30 when the norm is violated. This would solve the third-order free-riding problem. If actor B were to receive such a $40 reward, he would be happy to punish actor A because doing so only costs $35. This would solve the second-order free-riding problem. In anticipation of receiving the $35 punishment from actor B, it is no longer in Actor A's interest to violate the norm and obtain $30. This solves the first-order free-riding problem and leads to the norm being observed.

[3] For simplicity of exposition, the diagram only shows costs of norm violation borne by actor C and costs and rewards when actor C rewards actor B for punishing actor A. The scenario is symmetric for costs of norm violation borne by actor B.

Diagram 1

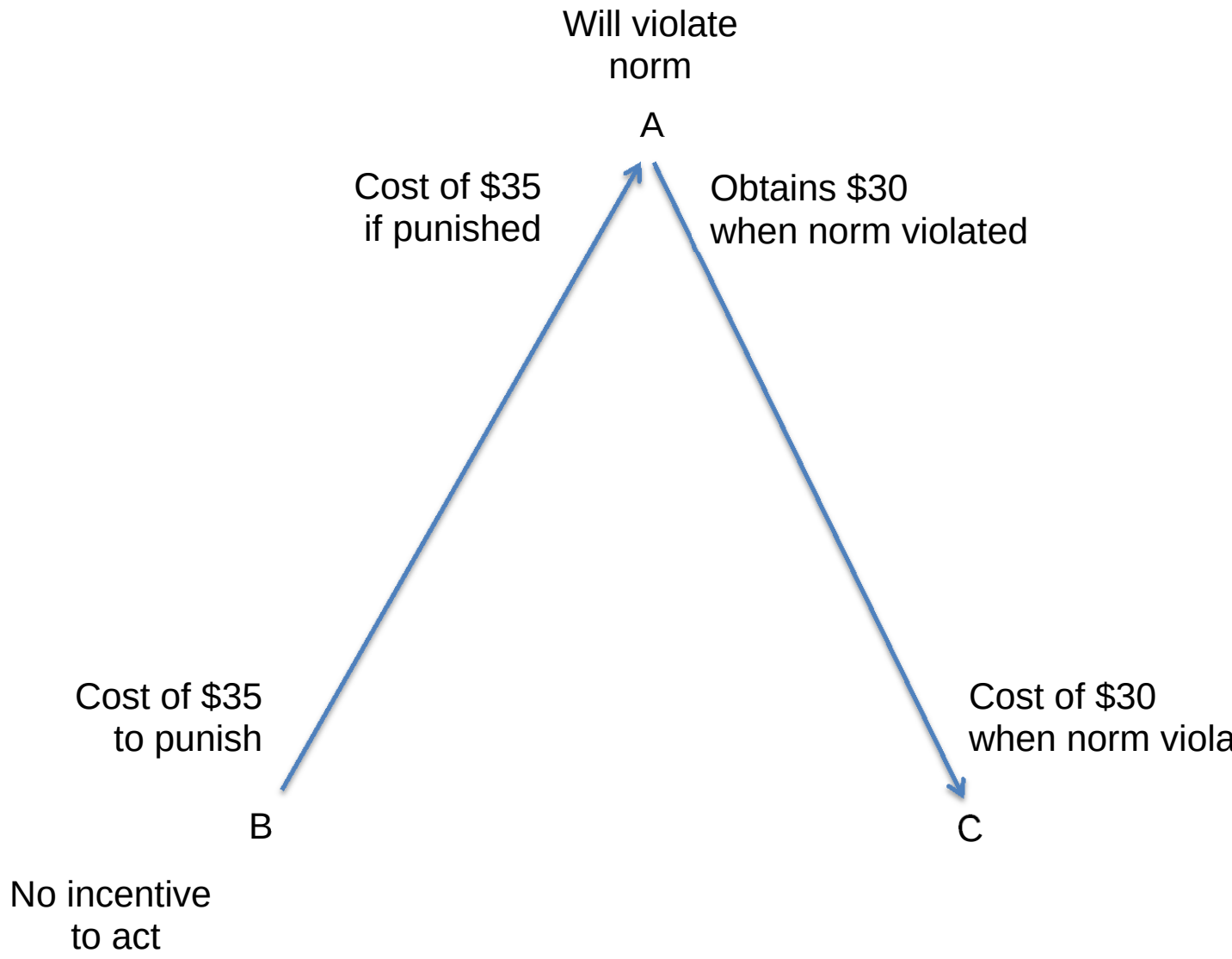


Density and the solution to the third-order free-rider problem

Within this framework, it is easy to understand the role of network structures and in particular the role of network density. Specifically, consider what would happen if there was no social relationship between B and C, such that C could no longer reward B for the act of punishing A's norm, as shown in **Diagram 2**. Put simply, without C being able to compensate B, the second-

order free-rider problem cannot be solved. As a consequence, actor A will not be punished in the event of norm violation, which will lead actor A to violate norms.

Diagram 2



This simple reasoning leads us to three pairs of hypotheses, one for each stage of the process. We will start with the final outcome of norm violations. Consistent with the discussion above, according to which actor A should engage in fewer norm violations in a high-density network, we argue that:

Hypothesis 1a: The higher an actor's network density, the less likely he or she is to violate a norm.

Since actor A should violate norms less frequently when density is high, actor C should experience fewer norm violations too. Hence:

Hypothesis 1b: The higher an actor's network density, the less likely he or she is to experience a norm violation.

For norm violations to occur less frequently under conditions of high density, it is necessary for punishments to occur more frequently. In the example above, actor B administered such punishments in anticipation of rewards from C. Because actor B is more likely to punish violations under conditions of high network density, we hypothesize:

Hypothesis 2a: The higher an actor's network density, the more likely he or she is to punish a norm violation.

If in high-density networks actor B punishes actor A more frequently for inflicting norm violations against C, it should also be the case that actor C experiences more of punishments of actor A by actor B. Hence:

Hypothesis 2b: The higher an actor's network density, the more likely others are to punish norm violators on his behalf.

Finally, for the entire mechanism to function, it should be the case that in high-density networks actor C rewards actor B more frequently for punishing actor A. Thus:

Hypothesis 3a: The higher an actor's network density, the more likely he or she is to reward those who punish norm violations.

Since actor C is more likely to reward actor B under conditions of high density, it should also be the case that actor B obtains more rewards for punishing others under such conditions. Hence:

Hypothesis 3b: The higher an actor's network density, the more likely he or she is to be rewarded for punishing a norm violation.

Commitment to a social system

Thus far we have assumed that actors participate in a social system whether or not norm violations occur and whether or not violators are punished. In reality, however, actors can leave social systems to join others that will give them greater benefits. Such defections are particularly likely when actors are not heavily dependent on the social system and when they have easy access to alternative social systems to meet their exchange needs. We assume that, in contemplating such a move, actors compare the utility they derive from the current system to the expected utility of joining another one. When actors experience norm violations, particularly violations that go unpunished, they experience negative utility. The benefits of staying in the current social system thus decline in comparison to the next best alternative, making actors more likely to leave. This reasoning leads us to the following hypothesis:

Hypothesis 4a: An actor who experiences a norm violation whose perpetrator is not punished is less likely to continue participating in the social system.

Similar reasoning applies when an actor experiences a norm violation aimed at him or her, and chooses to punish the violator personally. Though the punisher may obtain some intrinsic benefits from doing so, he or she still incurs the costs of norm violation. As before, the benefits of staying in the current social system decline as compared to the next-best alternative, and he or she is more likely to leave the current social system. As a consequence, we hypothesize:

Hypothesis 4b: An actor who experiences a targeted norm violation and personally punishes the violator is less likely to continue participating in the social system.

The same reasoning leads us to the opposite prediction when a norm violation is met with a third-party punishment. In this case, the target of the norm violation suffers its cost, but that cost is then offset by the benefit of seeing the offender punished without having to incur the cost of punishment. As a consequence, the target will end up at least as well off as if no norm violation had occurred. Furthermore, the target now knows that in this social system similar norm violations will meet with third-party punishments in the future. Consequently, when comparing the current social system to another social system in which norm violations may or may not meet with punishments, he or she will be more likely to stay put. This reasoning leads us to hypothesize:

Hypothesis 4c: An actor who experiences a targeted norm violation that is punished by a third party is more likely to continue participating in the social system than if no norm violation had occurred.

The setting of the study

We test our hypotheses in the context of contributions to Wikipedia, the largest on-line user-contributed encyclopedia. Between its launch in 2001 and the end of 2007, Wikipedia attracted over 6 million registered editors; these contributors created over 2 million encyclopedic articles in English and over 7 million entries in 253 languages. The site has become the seventh most visited website in the world.

Wikipedia was built on an intuitive on-line platform called wiki software. Anyone with internet access could post a draft of an article as long as the topic was deemed suitable for an encyclopedia. With the exception of a few protected articles, anyone could also edit any article by adding new content or by editing or deleting existing content. When an editor saved such changes, the software created a new version of the article for everyone to see. The previous version was added to the article history page, together with the Wikipedia username of the editor who had saved it and the time and date when the version was saved.[4]

No one could act as the final arbiter of an article's content; a subsequent editor could edit any version further. To manage disagreements over content, Wikipedia asked editors to try to resolve differences of opinion via discussion. To ensure that such discussions did not interfere with article content, Wikipedia added a discussion page to each article. Wikipedia urged a focus on content and avoidance of *ad hominem* attacks and asked that editors act in good faith, signifying an intention to

[4] Since Wikipedia did not require editors to register a personal account to make most types of edits, some editors made changes anonymously. In these cases, the Internet Protocol address of the computer where the edits originated was recorded. Many editors did open accounts, however; doing so allowed them to compile a record of their contributions, and provided personal pages where they could introduce themselves and receive feedback from other editors.

help the project rather than hurt it, and assume that others act in good faith in the absence of clear evidence to the contrary.

Wikipedia rules required articles to be written from a neutral point of view, which meant that they should fairly represent all significant views on the topic that had been published in reliable sources. It also required that article content represent and cite publicly available research. An editor's contribution to an article was considered acceptable as long as he could furnish reliable sources that readers or other editors could easily check.[5] Despite these rules and earnest efforts to reach consensus, editors could not always reach a viable compromise. At that point, participants could have recourse to a formal dispute-resolution process.[6]

Many editors were happy with Wikipedia's editing process. "*I don't have a problem with people making changes to what I wrote as long as they, you know, have good reasons for making those changes*," said one editor we interviewed. "*You know, like making the article better.*"[7] Others found the process deeply troubling. "*There is no special treatment for experts or any way to bar anyone or group from changing the content*," said an editor who had stopped contributing. Indeed, Wikipedia's rules ensured that all editors were considered equal; no one's contributions were privileged by virtue of expertise in the field, advanced degrees or first-hand knowledge of the topic.[8]

---

[5] We use 'he' rather than 'he or she' because most Wikipedia contributors are men (see footnote 18 for details).

[6] The dispute-resolution process began with a "request for comment" from others, which allowed all editors to contribute their views on how the dispute should be resolved. Editors could also ask for assistance from a volunteer-run mediation committee or from volunteer Wikipedia editors who identified themselves as dispute-resolution specialists. If these measures proved insufficient, the mediation committee referred the case to the arbitration committee, staffed by 12–16 elected volunteers. That committee privately examined the entire record of all parties' conduct, paying particular attention to whether or not they had observed the good-faith rule. The committee then issued a public decision, which could ban an individual from engaging in particular behaviors or editing certain articles or from participating in Wikipedia in any fashion, either temporarily or permanently. The committee did not, however, rule on the "truth" of the underlying disagreement. By the end of 2006, the arbitration committee had ruled on over 100 cases.

[7] To collect interview data, we chose a random sample of editors from a list of current and past contributors available on Wikipedia. We contacted editors via e-mail and obtained a response rate of approximately 25 percent. We detected no response biases; the geographic and demographic profile of the editors we interviewed closely mirrors that of the entire Wikipedia population. At the request of editors, most interviews were undertaken via an instant-messaging program or free voice-over-IP programs. Interviews were analyzed using inductive methods to derive a theory of editor commitment, described in another paper by one of the authors. Quotes from the interviews are used here for illustrative purposes only.

[8] The only editors with special powers were a small group of administrators elected by consensus. These administrators were not employees of Wikimedia and did not enjoy special privileges when it came to content contributions or deciding on the value of others'

Norm violation: Undo

Wikipedia's open and democratic editing process made it possible for a kernel of an article to evolve very quickly into a full-fledged encyclopedia entry. It did, however, expose Wikipedia articles to acts of vandalism. Vandals—often unregistered editors—edited pages to add invective, deliberately replaced an entire article with invective or deleted article content altogether. To help editors recover valuable content after such acts of vandalism, Wikipedia attached an *undo* link to every version of the article on its history page (see **Figure 1**). By clicking that link, editors could swiftly undo the vandalized version of an article and replace it with the prior unaffected version. The vandalized version remained, however, in the history of article development. With this simple mechanism, Wikipedia editors were able to restore a page to its previous status as soon as an act of vandalism was detected.[9]

The undo link could also be used incorrectly. Although its use was intended solely to undo vandalism, some editors found it an easy way to assert their points of view on article content. By clicking on the undo link, an editor could remove all changes introduced by the previous editor and restore the prior version without bothering to re-edit the content or negotiate with the other editor. Use of the undo link in the absence of vandalism constituted one of the biggest normative violations on Wikipedia. It flouted the basic tenets of acting in good faith and assuming that others do so as well. Many editors we interviewed also described the violation as such: "*Imagine slogging over an article, trying to get all of the details right of something that happened 800 years ago, and then someone comes in and just erases you—no asking, no talking. . . . Poof, the content disappears! Can*

---

contributions. They were, however, given the power to delete Wikipedia pages if the editor community voted to do so, and to block editors whose actions were deemed antisocial.

[9] Wikipedia also allowed registered editors to sign up for a watchlist on any page, which alerted them promptly to any changes on pages they were watching.

you imagine anything more disrespectful?" Indeed, many editors who had left Wikipedia cited instances of their work having been undone as a key reason for leaving. One former Wikipedia editor said: "*I have a Ph.D. in South Asian musicology, so I really care that the Wikipedia entry reflects what we know about the topic. I spend a lot of time documenting everything on the appropriate pages, and then, bam. . . . Someone comes in and just undoes everything I have done. There is this one guy in particular does this to me all the time. So I try to talk some sense into him, but he won't talk. So I got really upset at all of this, and left.*"

The norm not to use the undo link (except when eradicating vandalism) has two characteristics that increase the likelihood that it will not be obeyed. First, as we described in Step 1 of the theory, this is an essential norm and therefore subject to first-order free-rider problems. All editors would prefer that the undo button not be used incorrectly and editors engage in the civil negotiation process over the article content. However, every one of them is tempted to use it to cheaply remove the content they disagree with. Second, this norm is at least partly disjoint in that those who are supposed to observe the norm (i.e. the editors of Wikipedia) are a smaller set than those who are the beneficiaries of norm compliance (e.g. the readers of Wikipedia who are not editors). This makes the free-rider problem ever more pronounced and again less likely that the norm will be observed. Given these conditions, any evidence of norm compliance should be seen as a conservative test of the underlying theory.

With these considerations in mind, we will treat use of the undo link as a norm violation (unless the undo removes profanity or reinstates an article after the bulk of its content has been removed). We will treat the editor who clicked the undo link as a norm violator. Consistent with **Hypothesis 1a** we expect that an editor embedded in a dense network will be less likely to undo an

article version saved by someone else. Furthermore, our discussion indicates that all editors are affected by this violation, but the main victim of the violation is the editor whose version was undone. After all, he put the effort to contribute the content and it is his content that was removed. For this reason we will designate the editor whose version was undone as the main victim of a norm violation. Consistent with **Hypothesis 1b** we expect that an editor embedded in a dense network will be less likely to experience an undo of an edit he saved.

Norm punishment: Revert of undo[10]

Because use of the undo link is readily apparent in an article's history, the author of an undone version and other editors will know that such a norm violation has occurred. Some editors ignore the undo and address the offending editor in good faith; others retaliate by clicking on the undo link themselves. This action, which undoes the previous undo and restores the prior version of the article, is known as *reverting the undo*. Because it conveys disrespect for the perpetrator of the first undo, that editor may respond with another undo, which may in turn be followed by another revert. Such skirmishes are known as "revert wars." To prevent them, Wikipedia has instituted a three-revert rule stipulating that no user can undertake more than three reverts on a given page within a twenty-four-hour period; violators are barred from making any changes to Wikipedia for a specified interval.

Many editors deal with undo actions on their own, but other editors and administrators can also step in to remind the offending editor that his or her actions are inappropriate. These reminders can take the form of a chastizing note posted on the personal talk page of the editor in question; alternatively, a third-party editor can express disapproval more actively by reverting the undo. Like the original undo, which sends a public signal of disrespect, a revert of undo by an editor who is not

[10] On Wikipedia, the terms *revert* and *undo* are often used interchangeably. (See, for example, http://en.wikipedia.org/wiki/Help:Reverting#Undo). To prevent confusion, we will refer to the initial act as an *undo* and the act of undoing the undo as a *revert of undo*.

the author of the undone version sends a public signal of condemnation of the undo act. It signals clearly that a third-party editor, uninvolved in the dispute, believes that the original undo was unjustified and that its perpetrator violated a social norm and should be punished.

The punishment of an undo of a revert has three characteristics that make it less likely to occur. First, as we argued in Step 2 of the theory part of the paper, eliciting norm compliance through punishments rather than rewards makes it less likely to occur. Second, the act of a revert is individual rather than group effort. Again, as we argued in Step 2, this will make a punishment less likely to occur. Finally, using a revert of an undo gives us an opportunity to observe punishment of norm violation by an unaffected third party. As we suggested in Step 2 of the theory, such third-party punishments are particularly unlikely. Taken together, these three conditions imply that punishments through reverts are unlikely to occur, suggesting that we offer a conservative test of the theory.

With these considerations in mind, we will treat a revert of undo as a punishment of a norm violation. We will consider the editor who reverted the undo as the punisher, and the editor whose version was reverted as the punished actor.[11] Consistent with **Hypothesis 2a** we expect that an editor embedded in a dense network will be more likely to revert an undo of an article version saved by another editor. Consistent with **Hypothesis 2b** we expect that an editor embedded in a dense network will be more likely to experience other editors revert an undo of an article version saved by that editor.

Rewards for punishing norm violators

[11] An editor other than the author of the undone version who undertakes a revert of undo may derive direct benefits from doing so if he or she cares about the quality of the article. If this is the case, the norm that the editor is enforcing is conjoint in nature. In the results part of this paper, we will distinguish between situations in which the reverter cares or does not care about the quality of the article and show that our results hold in both situations (see footnote 28).

Our interviews revealed that editors greet reverts with substantial gratitude. One commented: "*People undo my work. It does not happen all that often, but more often than I would like. And then before I know it happened, someone will come to my rescue and revert the undo without even telling me. It's only later that I find what happened when I look at the article history. It's sometimes people that worked with me on that article . . . but you know what's most interesting? . . . It's also people who worked with me on other stuff . . . meaning they are kinda looking out for me! I would sometimes shoot them a note to say thank you. I would also definitely look out for them in the future to see if someone undoes their work and when that happens I would revert that… you know... as a way to say thank you for what they did for me.*"

This comment and others we collected along similar lines suggest that editors who revert undos receive rewards from victims of undos. Such rewards can take the form of written expressions of thanks or reciprocal reverts of undos. For the purposes of our paper we chose to use reciprocal reverts of undos as a measure of rewards for punishing norm violators.[12] We thus expect that editors who revert an undo will be rewarded in the future when third parties revert undos of their work. Specifically, consistent with **Hypothesis 3a** we expect that an editor is more likely to revert undos of article versions saved by other editors who themselves reverted other undos, and this effect is particularly large when the editor is embedded in a dense network. Furthermore, consistent with **Hypothesis 3b** we expect that an editor embedded in dense network will be more likely to experience other editors revert an undo of an article version saved by that editor if that editor has reverted other undos. We expect this effect to be particularly large when the editor is embedded in a dense network.[13]

[12] Alternatively, we could have used a measure of frequency with which editors are rewarded by obtaining a private or public thank you message. We chose not to use this measure, as it is very difficult to collect reliably.

[13] The use of reciprocal reverts of undos as a measure of rewards for punishing norm violators provides us with a conservative test of Coleman's mechanism. As we explained above, the mechanism works most powerfully when the reward for punishment is cheaper to supply than the punishment itself. In our case, rewards for punishment are captured by reciprocal reverts, and punishments are

Data

To test these hypotheses in the context of Wikipedia, we obtained a dataset from the Wikimedia Foundation, the parent of Wikipedia, by downloading it from http://download.wikimedia.org. The dataset contains every version of every article contributed to the English-language Wikipedia site between January 2001 and October 2006.[14] For every article version, the dataset provides the time and date it was saved, the Wikipedia username (or Internet Protocol address) of the editor who saved it, and the version length in bytes. Having parsed the data, we wrote an algorithm in MATLAB, described in **Appendix A**, to help us identify counter-normative undos (i.e. excluding those that undid acts of vandalism) and reverts of undos.

Having run the algorithm across all articles in the dataset, we compared the resulting statistics to Wikipedia statistics and to those reported in other papers that tried to identify acts of undo and reverts of undo. The aggregate rates of undo and revert of undo identified by our algorithm are very similar to those reported in related work – roughly 7% of edits are undos or reverts of undos (Anthony, Smith, and Williamson 2009; Buriol, Castillo, Donato, Leonardi, and Millozzi 2006; Kittur, Suh, Pendleton, and Chi 2007).

Dependent variables

Armed with classifications of various sequences of article versions, we constructed the dependent variables needed to test our hypotheses. We did so by aggregating the occurrence of various norm

---

captured by reverts. Because the cost of undertaking a revert is similar to undertaking a reciprocal revert, Coleman's mechanism is likely to be weak. This makes it harder for us to detect evidence in support of that mechanism.

[14] The dataset also contains a complete record of discussion and talk pages, articles containing lists of other articles, and placeholder articles that merely redirect users to other pages. We exclude these auxiliary pages and analyze only the encyclopedia articles. The dataset does not include articles deleted prior to October 2006. This poses a potential problem, in that editors could have engaged in undo or revert actions on these articles, but a significant proportion of them were deleted because they contained very little content, and by implication generated little editing activity. Thus limiting ourselves to surviving articles does not substantially compromise our ability to detect acts of undo and revert of undo.

violations and punishments over a month-long period (i.e. t = one month).[15] First, to capture the extent to which a given editor violated norms on Wikipedia, we constructed a variable *Number of Times Editor i Undid Others*$_{it}$ equal to the number of instances when editor *i* undid any article version during time *t*. [16] We use this dependent variable in tests of **Hypothesis 1a**. To capture the extent to which a given editor experienced norm violations, we constructed a variable *Number of Times Editor i Was Undone*$_{it}$ equal to the number of instances when editor *i*'s article edits were undone by other editors during time *t*.[17] We use this dependent variable in tests of **Hypothesis 1b**.

To capture the extent to which a given editor punished others for violating the undo norm, we constructed two variables. To test **Hypothesis 2a**, we constructed a variable *Number of Times Editor i Reverted Others*$_{it}$ equal to the number of instances during time *t* when editor *i* reverted an undo of a version that another editor *j* had saved. To test **Hypothesis 3a**, we constructed a variable *Number of Times Editor i Reverted Others Who Reverted*$_{it}$ equal to the number of instances during time *t* when editor *i* reverted an undo of a version that another editor *j* had saved, as long as editor *j* had previously reverted another undo during time period *t*.

To capture the extent to which an editor *i* experienced others' punishing norm violations, we constructed three independent variables: (1) *Number of Times Editor i Was Undone Followed by No Revert*$_{it}$ equal to the number of instances during time *t* when editor *i*'s article edits were undone and received no reverts, (2) *Number of Times Editor i Was Undone Followed by Editor i Reverts Undo*$_{it}$ equal to the number of instances during time *t* when editor *i*'s article edits were undone by others and then reverted by the focal editor *i*, and (3) *Number of Times Editor i Was Undone Followed by*

[15] We also constructed the variables in two-week intervals, which increased the number of observations. Analyses using these variables produced equivalent results for the density measure across the models and yielded higher statistical significance. We thus report the more conservative results.

[16] We also constructed this measure using (1) the total number of edits, rather than instances, that were undone by editor *i*, and (2) an indicator variable that took the value of 1 if *Number of Times Editor i Undid Others*$_{it}$ was greater than 0, and zero otherwise. The results are not sensitive to how we calculated this measure.

[17] This count does not include acts of undo by self. We also constructed this measure using (1) the total number of edits, rather than instances, by editor *i* that were undone, (2) an indicator variable that took the value of 1 if *Number of Times Editor i Was Undone*$_{it}$ was greater than 0, and zero otherwise. The results are not sensitive to how we calculated this measure.

Another Editor Reverts Undo$_{it}$ equal to the number of instances during time *t* when editor *i*'s article edits were undone and then reverted by other editors. We use these dependent variables to test **Hypotheses 2b** and **3b**.

Finally, to capture the extent to which editors continue to contribute content to Wikipedia, we constructed a variable *At Least One Edit*$_{it}$ equal to 1 if editor *i* undertook at least one edit (excluding acts of undoing others' edits) during time *t* and zero otherwise. We use these dependent variables to test **Hypotheses 4a**, **4b** and **4c**.

Independent variables

Having defined our operationalization of the dependent variable, we now turn to the network of interactions among Wikipedia editors.[18] Editors rarely interact face-to-face, and most of their on-line interactions focus on content, rather than on socializing. As one editor said: "*I'm probably one of the editors who is more prone than others to behaviors such as engaging people and getting consensus for difficult changes that people are struggling over. . . . But even then, I keep my engagement focused on factual contributions and not really on on-line socializing.*" By working together, however, editors developed close social bonds. In the words of one editor: "*Even though you interact with people through text, it does tend to build community between editors. For example, I'm interested in taxation issues, and there are a lot of us interested in this topic. We have a really strong community, and I must say it keeps me coming back. If I was writing things completely in a vacuum, I would lose my interest.*" Not all editors were equally likely to experience such on-line

[18] Editors were volunteers, not employees of Wikipedia, and they did not receive direct monetary compensation for their contributions. According to Wikipedia's own surveys, over 86 percent identified themselves as male, and 70 percent reported being single. One-quarter were under 18 years old, one-quarter were between 18 and 22, one-quarter were between 23 and 30 and the remaining 25 percent were between 31 and 85. About one-third named a high-school diploma as their highest degree; 30 percent had an undergraduate degree and less than 20 percent had a master's degree or Ph.D. The same survey revealed wide variation in editors' motivations to contribute. "I liked the idea of sharing knowledge and want to contribute to it" and "I saw an error and wanted to fix it" were the two most frequently cited reasons for contributing. The least frequently cited reasons for contributing were a desire to make a reputation in the Wikipedia community, ambition to make money and fondness for mass collaboration.

relationships. One explained: "*It's not like there is this one big Wikipedia community. There are communities inside the community. Some are strong; some are weaker. My personal experience is that most of the time, editing Wikipedia, I am doing it on my own and don't often encounter the same editors repeatedly.*"

On the basis of these statements we chose to use prior interactions between editors on the same articles as our measure of relationships between editors. To capture these interactions we wrote another algorithm which coded editor *i* and editor *j* as both contributing to the same article *a* if editor *i* had contributed at least one edit (excluding undos and reverts) to article *a* during period *t*, and editor *j* had contributed to the same article *a* during the same time period.

Some articles on Wikipedia, such as those about the World Cup, George W. Bush and Jesus, attract as many as 5,000 editors. It is hard to make the case that these editors interact with each other on these articles; many edit without being aware of each other's existence. By contrast, contributors to articles with fewer total editors are keenly aware of each other's existence and describe the process of editing as interaction. We thus decided to include only articles with fewer than 25 registered editors in our calculation of relationships between editor *i* and editor *j*.[19] We then used these data to construct a symmetric editor-to-editor matrix R_t, whose elements, r_{ijt}, consist of the number of articles with fewer than 25 total registered editors during period *t* to which editors *i* and *j* both contributed during *t*.[20] On the basis of this matrix, we constructed $\mathbf{r}_{ijt}$ equal to 1 if $r_{ijt} > 0$, and zero otherwise.

[19] We tested the sensitivity of our results to this restriction and found that coefficient estimates on the variables we use to test our hypotheses are still in the expected direction, though the statistical significance of the estimates is substantially lower across almost all of the specifications.

[20] It is also possible to define R_{ijt} as the number of total edits editor *i* contributed to articles to which editor *j* also contributed. This approach makes R_{ijt} asymmetric, and thus makes the empirical analysis more complicated. It also tends to make the relationship of *i* to *j* strong if *i* made numerous edits to a particular article. For this reason, we report the simpler analysis. Auxiliary analyses using the simpler approach yielded similar results but with weaker statistical significance.

Using this definition we constructed a simple measure of density around editor *i*, we calculate the number of relationships between editors with whom editor *i* has co-edited, as represented by $\bar{r}_{ijt}$, and divide it by the number of possible relationships between editors with whom editor *i* has co-edited. Using *q* as the total number of editors in the dataset at time *t*, we defined this measure as: [21]

$$Density_{it} = \frac{\sum_{m=1}^{m=q}\left(\sum_{n=1}^{n=q}\bar{r}_{im}\bar{r}_{mn}\right)}{\left(\sum_{m=1}^{m=q}\bar{r}_{im}\right)\left[\left(\sum_{m=1}^{m=q}\bar{r}_{im}\right)-1\right]}$$

Control Variables

Since density measures depend on the number of different articles editor *i* has edited, as well as the number of other editors who co-edited those articles, we include them as controls. First, we calculated a measure of *Number of Articles Edited*$_{it}$, equal to the number of articles that editor *i* edited during time *t*. Second, we captured the extent to which editor *i* edited the same articles repeatedly by constructing *Percentage of Articles Editor i Edited More than Twice*$_{it}$, equal to the

---

[21] The results we report below are based on density measures using only the existence of a relationship, $\bar{r}_{ijt}$, rather than its strength r_{ijt}. In auxiliary analyses, we develop alternative measures using relationship strength and generate very similar results. We report results based on simpler variable definitions. We also test for one other specification of the density measure to protect ourselves from the following situation: three editors, *i*, *j*, and *k*, work on the same article; it is the only article they work on. If this is the case, *i* and *j*, *j* and *k* and *k* and *i* will each have a tie to each other and to no one else, and as such *i*, *j*, and *k* will be surrounded by a perfectly dense network. Such an environment would be likely to generate very few undos, and those that occurred would be quickly reverted by one of the three highly committed editors. As a consequence, we would observe a relationship among density, a low incidence of undos and a high incidence of reverts. This empirical observation would probably be an artifact of having three editors deeply committed to the article; it would have little to do with the mechanism we seek to test here. To protect ourselves from such a statistical artifact, we calculate another measure of density that excludes participation in the same article by *j* and *k* when *i* is present, given by $\tilde{r}_{ijkt}$. The formula is given by:

$$Density_{it} = \frac{\sum_{j=1}^{j=q}\left(\sum_{k=1}^{k=q}\left(\bar{r}_{ijt}\bar{r}_{jkt}-\tilde{r}_{ijkt}\right)\right)}{\left(\sum_{j=1}^{j=q}\bar{r}_{ijt}\right)\left[\left(\sum_{j=1}^{j=q}\bar{r}_{ijt}\right)-1\right]}$$

This is a very conservative estimate of triadic relationship between *i*, *j*, and *k,* because in this specification it is possible that editor *i* has co-edited with *j* on one article and with *k* on another article but is unaware that *j* and *k* co-edited another article together. If *i* is unaware of this relationship, he might fail to act in the manner described by the theory. Results for this specification are in the same direction as those for the variable defined in the body of the text, but the statistical significance of the results is often weaker.

number of articles with two or more edits by editor *i* during time *t* by the number of articles editor *i* edited during time *t*.

Third, we included *Network Size*$_{it,}$ equal to the log of the total number of editors across all of the articles that editor *i* edited during time *t*. Fourth, we constructed variables *NetworkSize0*$_{it}$ and *NetworkSize1*$_{it}$ to reflect the fact that when variable *Network Size*$_{it}$ takes the values of zero and one, it is impossible to define measures of density. In such situations, we assigned a value of zero to *Network Density*$_{it}$. To differentiate this zero from editors' actual scores of zero, we assigned a value of one to *Network Size0*$_{it}$ when editor *i* edited other articles with no other editors during time *t*. Similarly, we assigned a value of one to *Network Size1*$_{it}$ when editor *i* edited other articles with only one other editor during time *t*.

Finally, we included other measures that are not necessarily directly correlated with density but that can influence the extent to which editor *i* experiences norm violations or norm restitutions. First, we included *Cumulative Edits*$_{it}$, equal to the log of the cumulative number of edits by editor *i* prior to *t*, as well as the square of that number. Second, we constructed *Months since Signup*$_{it}$ equal to the number of months since editor *i* first registered on Wikipedia. Finally, we constructed month dummy variables, *Time Period Dummies*$_{t}$, to control for temporal heterogeneity in norm violations, restitutions and project involvements.

Risk set

It was our intention to examine undo and revert-of-undo actions by all registered editors in our dataset.[22] Our preliminary analyses revealed, however, that although there are over 600,000 editors in our dataset, almost 125,000 of them contributed only one edit, another 50,000 edited only twice

[22] It is possible to contribute to Wikipedia without registration, in which the edit is recorded together with the Internet Protocol address of the computer from which the change was made. Since it is possible that many different editors used the same computer to make changes (e.g. university library), we chose to exclude edits by unregistered editors and only focused on registered ones.

and roughly another 30,000 contributed no more than three edits. Our interviews revealed that such editors are unlikely to be familiar with Wikipedia rules, and are therefore more likely to commit editing mistakes, e.g., to introduce a controversial point of view to the article without checking the article's talk page, where other editors may already have discussed how to handle this point of view. Edits by such inexperienced editors are often undone by existing editors without subsequent reverts of undo.

This dynamic is problematic for our analysis, because inexperienced editors have not had an opportunity to develop a dense network. Thus we are more likely to observe a positive relationship between low network density, a high incidence of undo and a low incidence of reverts of undo. Though this empirical observation is consistent with our predictions, it is not generated by the mechanism we want to test. To provide a more conservative test of our hypotheses, we chose to include an editors in the risk set only after he had contributed 25 edits, thus restricting our sample to 36,194 editors. Though this sample represents only a small subset of all Wikipedia editors, and thus raises issues of selection biases, it is reassuring to know that the editors in our sample contributed 70 percent of all edits.[23]

We then examined the timing of edits and found that as many as 40 percent of editors who edit one year do not do so the next year, suggesting that year-long data panels might be sufficient to capture most of the variation. We also found that 2005 was the most prolific complete year in our sample. That year alone witnessed the entry of more than 125 percent as many editors as there had been between 2001 and 2004. These editors contributed four times as many edits as they had

[23] In choosing the cutoff point, we considered the following tradeoff. An increase in the cutoff point reduced the number of editors in the sample and thus restricted the percentage of all edits under consideration. On the other hand, it increased editors' familiarity with Wikipedia's norms and made it less likely that we would be unable to define our density variables. (This consideration applied in particular to editors who edited articles singlehandedly without others' contributions.) We found that increasing the cutoff point to a minimum of 30 edits led to a substantial decrease in the percentage of all edits considered but had very little impact on our ability to define the density variable. On the other hand, lowering the cutoff point to a minimum of 20 edits had a much smaller effect on the percentage of total edits considered but a large effect on our ability to define the density variable. As a consequence, we chose 25 as our cutoff point.

between 2001 and 2004. Finally, in 2005, the rate of undos and reverts of undos increased more than twofold compared to the period between 2001 and 2004. For these reasons, we chose to focus on the time period between January and December 2005. Table 2 summarizes descriptive statistics for the dataset analyzed.[24]

Models

To test **Hypotheses 1–3**, we used random-effect negative binomial models, which we constructed as follows. Following Hausman, Hall and Griliches (1984), we assumed that the value of the dependent variable for editor i at time t followed a Poisson distribution: $DepVar_{it} \sim Poisson(\gamma_{it})$ where $\gamma_{it} \sim gamma(\lambda_{it}, \delta_i)$. The parameters of the gamma distribution are $\lambda_{it} = exp\ (\beta x_{it} + \varepsilon_{it})$, with vector x_{it} capturing independent variables for editor i at time t, and β is a set of parameters to be estimated. Parameter δ_i captures dispersion (i.e., variance divided by the mean) for editor i. On the basis of these assumptions, we define Γ as a gamma function, and construct a basic negative binomial model:

$$\Pr\left(DepVar_{it} = depvar_{it} \,|\, x_{it}, \delta_i\right) = \left(\frac{1}{1+\delta_i}\right)^{\lambda_{it}} \left(\frac{\delta}{1+\delta_i}\right)^{depvar_{it}} \frac{\Gamma\left(\lambda_{it} + depvar_{it}\right)}{\Gamma\left(\lambda_{it}\right)\Gamma\left(depvar_{it}+1\right)}$$

This model assumes, however, that the dispersion is constant across editors. To derive a random-effect negative binomial model, we allow δ_i to vary randomly across editors and assume $1/(1+\delta_i) \sim beta(r,s)$. Using f for the probability density function of δ_i we get the joint probability of the dependent variable for editor i at time t:[25]

[24] To test the robustness of our results, we re-ran our models for editors with more than 25 edits during 2004. The coefficient estimates on density variables remain in the predicted direction, but given the smaller frequency of undos and reverts, the statistical significance of the results is less robust.

[25] We also constructed fixed effect negative binomial models as described by Allison and Waterman (2002) to remove all types of time invariant unobserved heterogeneity for editor i. Such estimations yield coefficient estimates on the main variables of interest that are directionally similar to those of random effects. However, the fixed effect estimation procedure assumes that the individual fixed effect is related to the individual dispersion parameter δ_i through a specific functional form, e.g. fixed effect is the logarithm of the dispersion parameter (Hausman, Hall, and Griliches 1984). Guimaraes (2008) developed a method to test this assumption which we undertook on our data. We found that the test is not met, implying that the fixed effect negative binomial model might not perform

$$\Pr\left(DepVar_{i1} = depvar_{i1},...,DepVar_{in} = depvar_{in} \left| x_{i1},..,x_{in} \right.\right) =$$

$$= \int_0^{\infty} \prod_{i=1}^{n} \Pr\left(DepVar_{it} = depvar_{it} \left| x_{it}, \delta_i \right.\right) f\left(\delta_i\right) d\delta_i =$$

$$= \frac{\Gamma\left(r+s\right)\Gamma\left(r+\sum_{t=1}^{n}\lambda_{it}\right)\Gamma\left(s+\sum_{t=1}^{n}depvar_{it}\right)}{\Gamma\left(r\right)\Gamma\left(s\right)\Gamma\left(r+s+\sum_{t=1}^{n}\lambda_{it}+\sum_{t=1}^{n}depvar_{it}\right)} \prod_{i=1}^{n} \frac{\Gamma\left(\lambda_{it}+depvar_{it}\right)}{\Gamma\left(\lambda_{it}\right)\Gamma\left(depvar_{it}+1\right)}$$

Results

Step 1: Violating norms

With this specification, we estimated the likelihood that editor *i* violates norms on Wikipedia during time *t* by measuring the number of acts of undo undertaken by *i* against other editors, $DepVar_{it}$ = *Number of Times Editor i Undid Others*$_{it}$. It is possible that *Number of Times Editor i Undid Others*$_{it}$ takes on a value of zero if editor *i* does not engage in any undo actions, whether absent from Wikipedia or fully engaged in the project. In order to focus only on situations when editor *i* performs no undo actions while fully engaged in Wikipedia, we want to control for situations in which he is absent from Wikipedia. We do so in a number of ways. First, we eliminate from the risk set for editor *i* all time periods *t* during which he did not contribute any edits. Second, we retain all time periods but include a dummy that takes the value of one when editor *i* contributed no edits during time *t*. Finally, we also use Heckman-like correction (Heckman 1979) and estimate the likelihood that editor *i* will contribute at least one edit during time *t* as a function of z_{it}, given by $Edits_{it} = z_{it}\psi + \omega_{it}$ where z_{it} includes selection-independent variables. Logit estimates of this equation are then used to derive the inverse Mills' ratio, given by $InverseMills_{it} = \phi(z_{it}\hat{\psi}) / \Phi(z_{it}\hat{\psi})$ where ϕ is probability density function, Φ is the cumulative normal density and $\hat{\psi}$ is the estimate of ψ. This ratio gives us

reliably here. As a consequence, we prefer to report results from the more reliable random effect negative binomial, taking solace in the fact that the results are directionally similar across the two types of models.

the probability that editor *i* contributes at least one edit during time *t* given what we know about his or her characteristics as an editor. We include $InverseMills_{it}$ as an independent variable in the estimations of the *Number of Times Editor i Undid Others*$_{it}$ model. All three methods yield the same results for the density measure, and for brevity we report only the uncorrected results and those with $InverseMills_{it}$.

Table 3 reports the results of these estimations. Consistent with **Hypothesis 1a**, we find that editors embedded in dense social networks are less likely to undo other editors' edits. This effect holds across all six models and is therefore robust to various specifications. As for control variables, we find that editors of articles that were not edited by anyone else were less likely to engage in acts of undo, and that those who co-edited with one other editor were no more likely to engage in such acts than those who co-edited with two editors. Beyond that, an increase in the number of co-editors led to an increase in the likelihood of performing an undo. Similarly, the total number of articles edited by editor *i*, as well as higher percentage of articles edited more than twice led to a higher incidence of engaging in an undo. Editors who had signed up a long time earlier were also more likely to perform undos, but that effect was offset by the negative effect of actually contributing edits to a project. Finally, editors who had had their own edits undone by others were more likely to undo others' edits.

To test **Hypothesis 1b**, we model the likelihood that editor *i* suffers a norm violation during time period *t*, $DepVar_{it}$ = *Number of Times Editor i was Undone*$_{it}$. Editor *i* may suffer no undos because he does not contribute to Wikipedia, or alternatively because no one undoes his or her edits even when they are numerous. In order to focus on the latter scenario, we again control for the possibility of the former in the three ways described above. Coefficient estimates on the density variable are in the same direction across all three methods, and for brevity we report only results

using the Heckman-like correction. **Table 4** reports the results of these estimations. Consistent with **Hypothesis 1b**, we find in Models 7–9 that editors embedded in dense social networks are less likely to suffer an undo.

To test the robustness of our results, we checked whether the model can predict the number of times editor *i* was undone during time *t*, contingent on editor *i* being undone at least once during that period. To do so, we estimated $\textit{Was Undone at Least Once}_{it} = y_{it}\mu + \tau_{it}$ where y_{it} includes a set of selection-independent variables, and then estimated $\textit{InverseMills}_{it} = \phi(y_{it}\hat{\mu}) / \Phi(y_{it}\hat{\mu})$ where ϕ is probability density function, Φ is the cumulative normal density and $\hat{\mu}$ is the estimate of μ. We then include that $\textit{InverseMills}_{it}$ estimate in the random-effect negative binomial regression of *Number of Times Editor i Was Undone*$_{it}$. We report these results in Models 10–12 and obtain results directionally similar results to those in Models 7– 9.[26]

Steps 2 and 3: Eliciting norm compliance and compensating those who elicit norm compliance

To test **Hypotheses 2a** and **3a**, we modeled the likelihood that editor *i* reverts an undo during time period *t*. We distinguish between two types of reverts of undo: (1) editor *i* steps in to revert the undo of an article version saved by another editor, as captured by $\textit{DepVar}_{it}$ = *Number of Times Editor i Reverted Others*$_{it}$, and (2) editor *i* steps in to revert an undo of an article version saved by another editor *j* who has at least once reverted an undo of another editor's work during (t-1), as captured by $\textit{DepVar}_{it}$ = *Number of Times Editor i Reverted Others Who Reverted*$_{it}$.

[26] We also find that editors of articles edited by no one else were less likely to suffer an undo, and that those who co-edited with one other editor were no more likely to engage in undos than those who co-edited with two editors (once the number of articles was controlled for; see Models 9 and 12). Beyond that, an increase in the number of co-editors led to an increase in the likelihood of experiencing an undo. Likewise, the total number of articles edited by editor *i*, as well as his or her focus on a small number of articles, led to a higher incidence experiencing an undo. Editors who had signed up a long time earlier were more likely to experience undos, but that effect was offset by the negative effect of actually contributing edits to the project. Finally, editors who undid the edits of others were more likely to experience undos themselves.

Table 5 reports the results of our estimations. In Models 13, 14 and 15, we examine the conditions under which, when another editor's article version is undone, editor *i* steps in to revert the undo. Across the three models, we find that editors embedded in high-density networks are not more likely to revert undos of another editor's work. This is inconsistent with **Hypothesis 2a**. In Models 16, 17 and 18, we examine the conditions under which editor *i* steps in to revert an undo of an article version saved by another editor *j* who had previously reverted an undo of another editor's work during (t-1), as captured by *Number of Times Editor i Reverted Others Who Reverted*$_{it}$. Across the three models, we find that editors embedded in high-density networks are more likely to engage in such behaviors and to reward those who had reverted others' work by reverting undos that affected them. This pattern of results supports **Hypothesis 3a**. Overall, this pattern of results suggests that Wikipedia editors in dense social networks do not blindly revert undos suffered by other editors. They only do that as a reward to others who engage in reverting undos for others.

To test **Hypotheses 2b** and **3b**, we modeled the likelihood that editor *i* experiences a revert of an undo during time period *t*. Here we use three dependent variables: (i) *DepVar*$_{it}$ = *Number of Times Editor i Was Undone Followed by No Revert*$_{it}$, (ii) *DepVar*$_{it}$ = *Number of Times Editor i Was Undone Followed by Editor i Reverts Undo*$_{it}$, and (iii) *DepVar*$_{it}$ = *Number of Times Editor i Was Undone Followed by Another Editor Reverts Undo*$_{it}$.[27] Editor *i* may of course experience no reverts simply because he did not suffer any undos, or because he suffered undos but no one reverted them. As before, we control for the former possibility in three ways. First, we exclude time periods *t* when editor *i* does not suffer any undos. Second, we include a dummy variable equal to one when editor *i* suffered any undos. Finally, we use Heckman-like correction, estimating the likelihood that editor *i* will suffer at least one undo during time *t*. The specification for this function was given in model 9.

[27] *Number of Times Editor i Was Undone Followed by Editor i Reverts Undo*$_{it}$ excludes situations in which editor *i* undid his own version and then reverted the undo.

Coefficient estimates on the density variable are in the same direction across all three methods, and for brevity we only report results using the Heckman-like correction.

Table 6 reports the results of our estimations. In Models 19, 20 and 21, we examine the conditions under which an article version saved by editor *i* was undone and followed by no revert. Consistent with **Hypothesis 2b**, across the three models we find that editors embedded in high-density networks are less likely to suffer an undo that is not followed by a revert. Consistent with **Hypothesis 3b**, across the three models we find that editors who have reverted undos of other editors' work are less likely to suffer an undo that is not followed by a revert and this effect is particularly strong if editor *i* is embedded in high-density network.

In Models 22, 23 and 24, we examine the conditions under which an article version saved by editor *i* was undone and editor *i* personally reverted the undo. Consistent with **Hypothesis 2b**, across the three models we find that editors embedded in high-density networks are less likely to revert an undo of their own work. Consistent with **Hypothesis 3b**, across the three models we find that editors who have reverted undos for other editors are less likely to revert an undo of their own work and this effect is particularly strong if editor *i* is embedded in a high-density network.

Finally, in Models 25, 26 and 27, we examine the conditions under which an article version saved by editor *i* was undone and then reverted by another editor. Consistent with **Hypothesis 2b**, across the three models we find that editors embedded in high-density networks are much more likely to experience an undo followed by a revert by another editor.[28] Consistent with **Hypothesis 3b**, across the three models we find that editors who have reverted undos for other editors are more

[28] We have also run auxiliary models in which we took the dependent variable from Models 19 to 21 and split it in two. First, we examine the conditions under which an article version saved by editor *i* was undone and then reverted by an editor with whom editor *i* has previously worked. Second, we examine the conditions under which an article version saved by editor *i* was undone and then reverted by another editor with whom editor *i* has not previously worked. Consistent with our expectations, the effect of density on the likelihood of a revert by another editor is higher if that editor has previously worked with editor *i*.

likely to experience an undo followed by a third-party revert and this effect is particularly strong when editor *i* is embedded in a high-density network.

Continued participation

To test **Hypotheses 4a**, **4b** and **4c**, which pertain to continued editor participation, we modeled the likelihood that editor *i* contributes at least one edit during time *t*. The dependent variable was coded one if the editor has contributed at least once during time *t* and zero otherwise. To estimate this model, we used a fixed-effects panel logistic, with joint probability function given by:

$$\rho_{it} = \frac{1}{1 + e^{-\beta x_{it}}}$$

Table 7 presents the results of Models 28, 29 and 30. Consistent with our expectations, we find across the three models that editors surrounded by dense network structures are more likely to continue contributing content to Wikipedia. In Model 28 we find that having one's contribution undone has on average no effect on the likelihood of continuing to write for Wikipedia. Model 29, however, reveals a great deal of variation in the effect of an undo on the likelihood of continued participation, depending on whether the undo was reverted and, if so, how. In contrast to the predictions of **Hypothesis 4a**, the results indicate that an undo left unreverted has no effect on the likelihood of continuing to contribute to Wikipedia.[29] Consistent with **Hypothesis 4b**, however, the results indicate that an undo personally reverted by the editor whose work was undone makes him or her less likely to continue contributing. Furthermore, consistent with **Hypothesis 4c,** the results indicate that an undo reverted by a third party makes the editor whose work was undone more likely to continue to contribute.

[29] We suspect that the lack of statistical significance occurs because some reverts go unnoticed by the editor, and thus are unlikely to have an effect on the editor's editing pattern.

For completeness, in Model 30 we also interact with the three variables associated with the three hypotheses with a measure of density around actor *i* at time *t*. We find that when editor *i* is surrounded by a dense network, the effect of unreverted undos remains the same. However, when editor *i* is surrounded by a dense network, the effect of undos of editor *i*'s work subsequently reverted by editor *i* is even more negative. This should not be surprising. An editor surrounded by a dense network is likely to expect that the network will revert the undo on his or her behalf. Failure of the network to do so, requiring the editor to step in and personally revert the undo, makes him or her more disappointed with the network and thus more likely to leave. In contrast, when editor *i* is surrounded by a dense network, the effect of undos of editor *i*'s work reverted by another editor is even greater. This should not be surprising. An editor surrounded by a dense network is apt to expect the network to revert the undo on his or her behalf. The expectation that the network surrounding him will continue to do that in the future makes the editor less likely to leave.

Limitations

The results we present provide overall support for our hypotheses, but they have some shortcomings. First, we do not directly measure relationships between individuals. Instead, we infer the existence of relationships by identifying who worked with whom on a given article. We believe, however, that this shortcoming does not undermine our results, and indeed that it makes our results conservative. Consider what would happen if we erroneously assumed that relationships exist when they do not—in other words, that a network of editors is dense when in reality it is not. We would expect these editors to behave in the manner described by the theory, but because there is no density between them they would not do so. As a consequence, we would be less likely to obtain the results we do. Conversely, we may have mistakenly assumed that relationships do not exist when in reality they do.

In this scenario, we would be underestimating the extent of density between editors—in other words, we would not expect these editors to behave as described by the theory, though in fact they do. This scenario too would make it less likely that we will find the results we do. Both of these measurement errors suggest that our results are fairly conservative estimates.

We also labor under the disadvantage of being unable to measure all types of norm violations. This would be a problem if, for example, editors in dense networks were less likely to undo edits, but more likely to violate norms on, say, the talk pages where editors discuss how an article should evolve. If this were the case, however, we would expect extensive spillovers, such that editors who violate norms on, say, talk pages would be more likely to experience retribution in the form of undos of their article versions. This scenario should lead to a positive relationship between density and the likelihood of experiencing undos, making it less likely that we will observe a negative relationship between the two. Thus the negative relationship we document should be seen as a conservative estimate.[30]

Finally, we are unable to capture all types of punishment of norm violations. For example, some editors who undid articles might have been punished via private e-mails. This scenario could present a problem for interpretation of our results in the following way. Suppose no real relationship exists between density and reverts of undos, but editors do tend to chastise other editors embedded in sparse networks via private communication, and those embedded in dense networks via public reverts of their undos. If this differential treatment existed, we would observe a relationship between density and reverts of undos even if it did not exist. It is very unlikely, however, that this scenario actually prevails. If anything, we would expect editors embedded in dense networks to be less likely than those with sparse networks to have their undos reverted (for fear of retaliation, say). Thus this

[30] Similar logic could be applied to an unobserved propensity of editors in high-density networks to experience norm violations on, say, talk pages. Again, to the extent that editors penalize such behavior via undos, we should observe a positive rather than negative association between density and undos.

potential bias makes the results we observe less rather than more likely. Finally, it is unlikely that we capture all types of rewards for those who punish norm violations. Such rewards can take the form of private thank-you e-mails, public thank-you entries on editors' private pages, and other expressions of gratitude. Once again, to the extent that such rewards are substitutes for reverts of undos, we should be less likely to observe the results we do.

Finally, there remain the issues of reverse causality and unobserved heterogeneity. With respect to reverse causality, it is possible for an individual editor to create a high-density network around himself or herself by introducing acquaintances to each other. In gratitude for such introductions, the acquaintances may in turn refrain from violating norms against the editor, or may punish those who do so. Though this scenario could occur, we have undertaken a number of steps to exclude it from the data. Suppose editor *i* works with editor *j* on article A, and with editor *k* on article B. Editor *i* may tell editor *k* about his work with editor *j* and invite him to join the two of them in working on article A, which editor *k* does. Because our definition of network density around editor *i* explicitly requires editors *j* and *k* to work together on a different article in which *i* does not participate, those two editors would then have to start editing another article, C. They would also have to attribute this new undertaking to editor *i*'s introduction, and in gratitude perform fewer undos or revert more undos affecting editor *i*. We doubt that such joint editing activity on article C would be attributed to editor *i*, and thus we do not believe that reverse causality is responsible for our results.

Concerns about unobserved heterogeneity can also lead one to argue that editors engaging in or suffering from fewer norm violations; engaging in or witnessing more punishments of norm violators; and rewarding those who punish norm violations, as well as getting rewarded for such acts, find themselves in this situation not because of density, but because of their unobserved

personal characteristics. A critic can then argue that these unobserved characteristics are correlated with editor's proclivity to form dense networks, which results in the empirical association between density and the six types of behaviors described above. We believe that these concerns are attenuated by the fact that auxiliary analyses we ran using fixed effect models (see footnote 25 and examine estimation procedure for models 28-30) generate similar pattern of results, imply that the time invariant unobserved characteristics cannot be held responsible for generating the results. The unobserved heterogeneity explanation of our results is thus limited only the time-varying unobserved effects.

Conclusions

Since the inception of the discipline, sociologists have examined the role of dense social relationships in various social phenomena. People surrounded by friends who are also each other's friends are thought to enjoy more social and economic support (Durkheim 1951; Uehara 1990). They are also believed to be less likely to commit or suffer norm violations. Coleman (1990) formalized this intuition and argued that high-density networks enable third parties to compensate norm enforcers for the expense of chastising norm violators. Such payments encourage actors to punish those who violate norms, in turn reducing the incidence of norm violation. Despite ubiquitous citations of Coleman's explanation, little empirical work has tested it convincingly. This is problematic; we do not know whether the mechanism is borne out in reality. If not, we may erroneously recommend that a network be made denser even if doing so will not improve norm enforcement. Our paper endeavors to address this issue by testing Coleman's mechanism in detail. We find substantial support for it, suggesting that increasing network density to elicit norm compliance is justified. Support for Coleman's mechanism alerts us to the importance of

punishments for norm violations and rewards for such punishments, and thus helps us design social systems in which norms are observed.

The fact that we found supporting evidence in the Wikipedia context highlights a number of conditions that promote the operation of Coleman's mechanism. On Wikipedia, for example, norm violations, punishments of norm violations and as rewards for punishing norm violators are all highly visible. Replicating these conditions in the design of a social system is critical; otherwise norm violations will remain undetected and therefore unpunished. Wikipedia's norms are also clearly articulated, making it easy to detect a violation and fairly difficult to claim that a norm violation occurred when it did not. It is also reasonably clear how to punish violators in ways that will elicit rewards from others. Without such clear specification of appropriate punishment, some actors may be afraid to administer it for fear of committing a violation themselves.

Understanding such conditions has important implications for related streams of the literature, such as the effort to link network density to performance. On the one hand, higher network density is believed to constrain the novelty and creativity of new ideas and solutions, and thus individual and collective performance (Burt 2005). On the other hand, higher network density is thought to enhance performance via a higher rate of norm compliance (Uzzi 1999). Numerous papers seek to address this tradeoff by pointing to sets of conditions under which one or the other effect is likely to be stronger, suggesting for example that performance will be higher in dense networks when tasks are collective and require everyone's cooperation (Ahuja 2000). Our results indicate that this positive association will hold only when mechanisms for punishment and reward of punishment are in place. Otherwise, dense networks will suffer all the shortcomings of constrained creativity without enjoying any of the benefits of higher norm compliance.

The theory and results we present here also inform our understanding of what makes social systems survive. Specifically, they underscore a tradeoff in designing a social system between maximum norm compliance and maximum longevity. Our results indicate that norm violations followed by punishments make people more committed to a social system than they would be if they had never experienced a norm violation. To the extent that very dense networks discourage norm violations, they also prevent actors from learning just how strong the community is. Even a small decrease in network density will increase the rate of norm violation, whose punishment will in turn promote greater commitment to the social system. These conclusions are similar in nature to those of Uzzi (1999), who found that intermediate levels of density promote the highest performance. Whereas Uzzi's quantitative findings pertain to performance, our paper quantitatively estimates commitment to a social system.

We hope that our paper will stimulate further research. We see substantial opportunities for further tests of Coleman's mechanism. Specifically, it would be helpful to document the conditions under which network density has no effect on norm enforcement. If Coleman's theory is correct, for example, it should be the case that when norm violations, punishments and rewards for norm punishments are hard to observe, density will have limited effect on these phenomena. Density should also have no effect on populations of individuals who derive sufficient intrinsic rewards for punishing those who violate norms. Furthermore, density will not lead to norm observance when rewards for punishment are very expensive to provide, such that a third-order free riding problem occurs. Finally, further opportunities exist to show that density may actually reduce the incidence of norm observance. This mechanism would be most likely if punishing friends who are also each other's friends were particularly costly. Demonstrating that the relationship between density and

norm observance critically depends on such factors would further lend credence to Coleman's theory.

We hope that future research will take advantage of the vast amounts of data on social interactions on the internet. This unprecedented opportunity for insight into human interactions makes it possible to offer unequivocal empirical support for many theories central to sociology. For example, a string of papers using e-mail data has convincingly shown that homophily, as distinct from other mechanisms, does indeed explain why actors with similar characteristics are more likely to form relationships with each other (Kossinets and Watts 2009; Menchik and Tian 2008). This paper too provides empirical support for a widely accepted mechanism. It is to be hoped that future papers will furnish unambiguous evidence for other widely cited social theories. We hope too that a new set of papers will take advantage of the fact that on-line environments make certain social mechanisms more salient, allowing for development of new theories. For example, Piskorski (2010) has shown that on-line social networks allow people to create an illusion of constant sociability, which they can then use to engage in other, often illegitimate, activities. Similarly, one can argue that such social platforms make others' patterns of social relationships public information, in turn illuminating opportunities for individuals to act as social brokers. Viewed as such, on-line environments can help us further our theories of brokerage. Other theory-development opportunities abound.

Appendix A

Vandalism

We first sought to discriminate undos of vandalism, which are legitimate actions, from undos of good edits, which are counternormative. To identify undos of vandalism, our sorted all the versions of a given article chronologically and analyzed each version starting with the oldest (see Table 1 for a sample article). The algorithm relied on the fact that an undo of vandalism creates a version of an article identical to a previous version. Because it would be too time-consuming to compare each version to all previous versions, we relied on the simple shortcut of comparing the lengths of successive versions.[31] That is, if no prior version of identical length existed, we concluded that the version in question could not be an undo of vandalism.[32] But if the algorithm found a previous version of the same article with an identical length, it examined all versions between the one in question and the previous version of the same length.[33] The algorithm then tested whether intermediate versions by the same editor were less than 10 percent of the size of the version in question. If so, we recoded all intervening edits as acts of vandalism and coded the version in question as an undo of vandalism.[34] In examining version 263 in Table 1, for example, the algorithm would discover that version 261 was the same length, and that version 262 was less than 90 percent as long. We would then code version 262 as vandalism and version 263 as an undo of vandalism.

[31] Another option would be to identify an undo of vandalism by examining the short notes that editors sometimes append when undoing an article version. But these notes are optional, and therefore unreliable.

[32] In some cases acts of vandalism take the form of very small changes to an article, such as inserting a vulgarity into the text. When the next editor undoes such an addition, this is technically an undo of vandalism. But because our algorithm identifies only large changes to articles as vandalism, such an act will be coded as an undone edit followed by an undo. The algorithm will thus underestimate the rate of vandalism and overestimate the rate of undone edits.

[33] If many versions of the same length were found, the algorithm would select only the most recent. For example, if the algorithm was currently analyzing version 263, both versions 242 and 261 would be identified as the same length as 263. The algorithm would then select version 261.

[34] It is conceivable that an act of vandalism that removed more than 90 percent of an article's content was followed by one or more versions and then by an undo that restored the article to its original state. To take this possibility into account, we assumed that the currently analyzed version was an undo of vandalism even if only one version existed with less than 90 percent of its content. We also tried coding the original edit that had removed 90 percent of article content as vandalism, the subsequent versions as regular edits, and the undo as an undo of vandalism. Given the rarity of such editing patterns, all coding schemes resulted in the same pattern of results.

Undo

We relied on the same logic to identify instances of an undo of a regular edit, which is normatively prohibited. This action creates a new version identical to a prior version but with no intervening vandalism edits.[35] If the algorithm detected such a pattern, the version in question was categorized as an undo and those between it and the prior version of the same length were designated undone edits.[36] This pattern can be seen in edits 264–267 in Table 1; version 267 is designated an undo and versions 265 and 266 are both categorized as undone edit.

The algorithm then sought to distinguish between a norm-violating undo and an acceptable scenario in which an editor created a new version of an article and then, dissatisfied with it, undid his or her own changes. To do so, the algorithm examined all intermediate versions between the version in question and the previous version of identical length. If all had been saved by the same editor, the algorithm qualified the undo as *undo by self* and undone edits as *undone edits by self*. This pattern can be seen in versions 267–270 in Table 1. If on the other hand the intermediate versions were saved by different editors, all acts of undo and all undone edits were coded as undertaken by other (see versions 264–267).[37]

[35] Our algorithm works only for the original implementation of the undo link, which restored a particular version of an article and removed all intermediate versions between it and the current version. For example, if the current version was 266 and an editor clicked the undo link next to version 264, the platform would create version 267 an exact replica of 264, and discard all changes introduced in versions 265 and 266. In late 2006, Wikipedia changed its software to allow editors to undo changes introduced in only one version while leaving others intact. Thus, in response to clicking the undo link next to version 264, the software would try to create a version 267 identical to version 266 but without changes introduced in version 264. This method retained the changes introduced in version 265. As a consequence, the resulting version was apt to differ in length from the undone version 264, and our algorithm would not work. As a consequence, we limited our analysis to the time period when the original undo regime was in operation.

[36] At this point, it would be possible to compare the texts of the two versions to ascertain that they are actually identical. However, such an exercise would require several terabytes of computer storage capacity and would lengthen the data analysis by many months. We thus use this speedier algorithm. Its biggest shortcoming is that it will assume that two versions are identical when they are merely of identical length, leading us to overstate the frequency of undo and revert-of-undo actions. To test the extent to which this is a problem, we chose 2,000 articles at random and identified every instance when our algorithm found two versions of equal length. We then wrote a short script to test whether the versions tagged as identical were in fact textually identical. Of pairs tagged as identical, 99.7 percent were found to be in fact identical. Such a high rate of correspondence makes us confident that our faster algorithm is relatively error-free.

[37] This algorithm assumed that all undone versions were undone by other, even if some of the intermediate versions were by the same editor who later undid the changes. To verify the robustness of our results, we re-ran the algorithm assuming that every undone version saved by the same editor who later undertook an undo was marked as undone by self and every undone version saved by a different editor was marked as undone by other. Designating undone edits by self and undone edits by other led to coefficient estimates that are in the same direction as those resulting from the algorithm in the main text. We report those from the main text.

Reverts of undo

Finally, we used similar logic one more time to identify reverts of undo. For a revert of undo to take place, an editor must first undo a prior version of the article, thus making that version an undone edit; another editor then reverts that undo and creates a new version identical to the version marked as undone edit. Thus we can identify a revert of undo by finding that a previous version of the same length has already been marked as an undone edit. If the algorithm found such a pattern, it would mark the currently analyzed version as a revert of undo and introduce no other changes. This pattern of edits can be seen in the sequence of versions 271–274 in Table 1. Version 271 was a regular edit, followed by edit 272, which was subsequently undone in version 273. Version 274, which reverted the undo in version 273, was identical to version 272. We will therefore code version 274 as a revert of undo.

Finally, the algorithm identified whether the revert of undo was undertaken by the same person who had authored the undone edit or by someone else. To do so, the algorithm compared the usernames of the editor who reverted the undo and the editor who saved the undone edit. If the names differed, as in the case of versions 272 and 274, the algorithm would tag the revert of undo as undertaken by other. If the two names were identical, as in versions 276 and 278, the algorithm would tag the revert of undo as undertaken by self.

Figure 1. Screenshot of a sample article history page

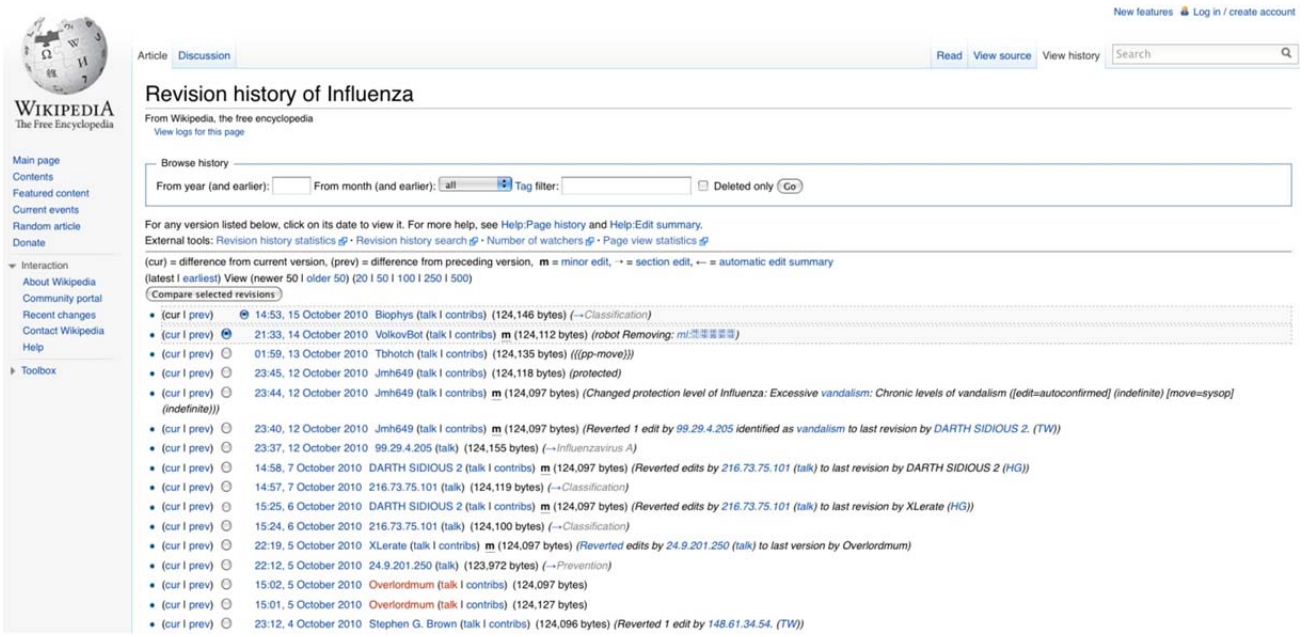


Table 1. Sample article history

Version	Date and time	Editor or IP	Text length in bytes	Final designation of edit type[38]	By
…	…	…	…	…	
242			94,399	Regular edit	
…	…	…	…	…	
261	Feb 4, 2005, 23:20	AndrewP	94,399	Regular edit	
262	Feb 4, 2005, 23:25	PettyCrime	10	Vandalism	
263	Feb 4, 2005, 23:27	DogEatDog	94,399	Undo of vandalism	
264	Feb 4, 2005, 23:15	DannyP	94,134	Regular edit	
265	Feb 5, 2005, 12:15	Angela	94,576	Undone edit	Other
266	Feb 5, 2005, 12:15	128.100.91.5	95,333	Undone edit	Other
267	Feb 5, 2005, 12:17	ZZTop	95,134	Undo	Other
268	Feb 4, 2005, 23:16	BriteLite	94,433	Undone edit	Self
269	Feb 4, 2005, 23:19	BriteLite	94,512	Undone edit	Self
270	Feb 5, 2005, 12:10	BriteLite	95,134	Undo	Self
271	Feb 5, 2005, 12:30	ZZTop	95,211	Regular edit	
272	Feb 5, 2005, 1:07	Angela	95,279	Undone edit	Other
273	Feb 5, 2005, 1:09	BlueHawk	95,211	Undo	Other
274	Feb 5, 2005, 1:10	DogEatDog	95,279	Revert of undo	Other
275	Feb 5, 2005, 1:12	Charlie	96,501	Regular edit	
276	Feb 5, 2005, 1:20	MustBeSerious	96,650	Undone edit	Other
277	Feb 5, 2005, 1:25	BlueHawk	96,501	Undo	Other
278	Feb 5, 2005, 1:27	MustBeSerious	96,650	Revert of undo	Self

[38] Because our algorithm moves forward in time, and we change designations by looking backward, acts initially coded in a particular way may be recoded when subsequent versions of the article are analyzed. Hence, we refer to it as the "final designation"

Table 2. Descriptive statistics

	Mean	Std. Dev.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1 Network Density$_{i,t}$	.12	.26															
2 Network Size$_{i,t}$	2.79	1.94	-.15														
3 Network Size0 $_{i,t}$	.20	.40	-.27	-.72													
4 Network Size1$_{i,t}$	.03	.17	-.10	-.19	-.09												
5 Number of Articles Edited$_{i,t}$	1.60	1.60	-.30	.45	-.59	-.13											
6 Percentage Articles Edited More than Twice$_{,t}$	.09	.19	.08	.15	-.24	.02	.13										
7 Number of Edits$_{i,t}$	1.70	1.69	-.21	-.61	-.32	-.09	.68	.12									
8 Cumulative Edits$_{i,t}$	3.62	2.12	-.28	-.30	-.27	-.12	.68	.07	.50								
9 Months since Signup$_{i,t}$	2.05	.95	-.06	.02	.08	-.03	.02	-.17	.01	.44							
10 Number of Times Editor *i* Undid Others$_{i,t}$	.35	.80	-.16	.50	-.19	-.06	.56	.07	.71	.49	.07						
11 Number of Times Editor *i* Was Undone$_{i,t}$	.41	2.67	-.06	.22	-.08	-.03	.26	.04	.24	.21	.02	.28					
12 Number of Times Editor *i* Was Undone Followed by No Revert$_{i,t}$	.35	2.32	-.06	.21	-.07	-.03	.25	.03	.22	.20	.01	.25	.54				
13 Number of Times Editor *i* Was Undone Followed by Editor *i* Reverts Undo$_{i,t}$	.04	.56	-.03	.11	-.04	-.01	.13	.04	.13	.12	.01	.21	.55	.36			
14 Number of Times Editor *i* Was Undone Followed by Another Editor Reverts Undo$_{i,t}$	.06	.41	-.05	.17	-.06	-.02	.19	.03	.18	.15	.02	.22	.60	.48	.30		
15 Number of Times Editor *i* Reverted Others$_{i,t}$	.23	2.66	-.02	.07	-.02	-.01	.07	.00	.08	.07	.02	.10	.03	.03	.01	.02	
16 Number of Times Editor *i* Reverted Others Who Reverted$_{i,t}$	.14	1.74	-.02	.06	-.02	-.01	.07	.00	.07	.07	.02	.09	.03	.02	.01	.02	.58

Table 3. Negative binomial random-effect estimates that editor *i* engaged in an undo during time *t* (Test of **Hypothesis 1a**)

	Number of Times Editor i Undid Others$_{it}$ Selection Equation: None			*Number of Times Editor i Undid Others*$_{it}$ Selection Equation: Editing		
Independent Variables	Model 1.	Model 2.	Model 3.	Model 4.	Model 5.	Model 6.
Network Density$_{i,t-1}$	-.42**	-.42**	-.37**	-.32**	-.22**	-.21**
	(.03)	(.03)	(.03)	(.03)	(.03)	(.03)
Network Size$_{i,t-1}$	.24**	.24**	.21**	.24**	.23**	.24**
	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)
Network Size0$_{i,t-1}$	-.30**	-.26**	-.23**	-.08*	.08*	.07*
	(.03)	(.03)	(.03)	(.03)	(.03)	(.03)
Network Size1$_{i,t-1}$	.02	.02	.02	.09*	.16**	.14**
	(.05)	(.05)	(.05)	(.05)	(.05)	(.05)
Number of Articles Edited$_{i,t-1}$	.14**	.13**	.18**	.14**	.20**	.17**
	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)
Percentage of Articles Edited More than Twice$_{i,t-1}$	.65**	.63**	.62**	.55**	.61**	.59**
	(.03)	(.03)	(.03)	(.03)	(.03)	(.03)
Cumulative Edits$_{i,t-1}$	.07**	.08**	.04**	-.01	-.22**	-.14**
	(.00)	(.01)	(.01)	(.01)	(.01)	(.01)
Months since Signup$_{i,t-1}$	—	-.02*	-.07**	—	.39**	.42**
		(.01)	(.01)		(.02)	(.02)
Number of Times Editor i was Undone$_{i,t-1}$	—	-.01*	-.01**	—	-.01	-.01
		(.00)	(.00)		(.01)	(.01)
Number of Edits$_{i,t}$*x10*	.05**	.07**	.06**	.07**	.12**	.08**
	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)
Inverse Mills Ratio$_{i,t}$	—	—	—	-.64**	-1.57**	-1.33**
				(.03)	(.06)	(.05)
Time Period Dummies$_t$	No	No	Yes	No	No	Yes
-Log-Likelihood	-206,772	-206,562	-205,720	-206,377	-206,139	-205,129
Degrees of Freedom	8	10	21	9	11	22
Wald χ^2	21,278	21,528	23,316	21,477	21,995	23,509

Note: Numbers in parentheses are standard errors. Constant was omitted. All χ^2 tests are based on a baseline model with no covariates. Results of sensitivity tests using generalized linear models with negative binomial link and grouped logits yield equivalent results. Selection equation predicting editing based on previous experience, tenure and month dummies was omitted from table. Resulting inverse Mills ratio was used as control in the outcome equation, as noted. Editors, i = 30,272; number of periods = 10; total number of observations = 212,317. (Not all editors started editing in time period 1.) *p < .05 **p < .01 ***p < .001 (two-tailed tests)

Table 4. Negative binomial random-effect estimates that editor *i* experienced an undo during time *t* (Test of **Hypothesis 1b**)

	Number of Times Editor i was Undone$_{it}$ Selection Equation: Edited			*Number of Times Editor i was Undone*$_{it}$ Selection Equation: Experienced undo		
Independent Variables	Model 7.	Model 8.	Model 9.	Model 10.	Model 11.	Model 12.
Network Density$_{it-1}$	-.18***	-.12**	-.27***	-.20***	-.25***	-.18***
	(.03)	(.04)	(.04)	(.03)	(.04)	(.04)
Network Size$_{it-1}$	.61***	.36***	.51***	.61***	.37***	.50***
	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)
Network Size0$_{it-1}$	—	-.32***	-.24***	—	-.58***	-.20***
	—	(.05)	(.05)	—	(.05)	(.05)
Network Size1$_{it-1}$	—	.06	.19***	—	-.03	.25***
	—	(.07)	(.07)	—	(.07)	(.07)
Number of Articles Edited$_{it-1}$	—	.23***	.24***	—	.20***	.21***
	—	(.01)	(.01)	—	(.01)	(.01)
Percentage of Articles Edited More than Twice$_{it-1}$	1.03***	.96***	.88***	1.07***	.97***	.83***
	(.03)	(.03)	(.03)	(.03)	(.03)	(.03)
Cumulative Edits$_{it-1}$	—	-.41***	-.36***	—	-.51***	-1.60***
	—	(.01)	(.02)	—	(.02)	(.05)
Months since Signup$_{it-1}$	—	.37***	.36***	—	.33***	1.37***
	—	(.02)	(.02)	—	(.02)	(.05)
Number of Times Editor i Undid Others$_{it-1}$	—	.42***	.44***	—	.40***	.44***
	—	(.01)	(.01)	—	(.01)	(.01)
Number of Edits$_{it}$ *x10*	.30***	.20***	.25***	.29***	.19***	.31***
	(.07)	(.09)	(.01)	(.07)	(.09)	(.01)
Inverse Mills Ratio$_{it}$	-.45***	-1.82***	-1.94***	-.24***	-1.51***	-5.90***
	(.03)	(.07)	(.09)	(.02)	(.08)	(.20)
Time Period Dummies$_t$	No	No	Yes	No	No	Yes
-Log-Likelihood	124,474	121,936	121,554	124,512	122,067	121,825
Degrees of Freedom	5	11	22	5	11	22
Wald χ^2	32,186	35,271	38,348	32,678	35,330	39,026

Note: Numbers in parentheses are standard errors. Constant was omitted. All χ^2 tests are based on a baseline model with no covariates. Selection equation predicting editing based on previous experience, tenure and month dummies was omitted from table. Resulting inverse Mills ratio was used as control in the outcome equation in Models 7, 8 and 9 as noted. The second selection equation was used to predict whether editor *i* experienced an undo during time *t* with the same independent variables. Resulting inverse Mills ratio was used as control in the outcome equation in Models 10, 11 and 12 to predict the number of undos conditional on experiencing at least one undo. Results of sensitivity tests using generalized linear models with negative binomial link and grouped logits yield equivalent results. Editors *i* = 30,272; periods *t* = 12; total number of observations = 212,317. (Not all editors started editing in time period 1.) *p < .05 **p < .01 ***p < .001 (two-tailed tests)

Table 5. Negative binomial random-effect estimates that editor *i* reverted an undo at time *t* (Test of **Hypotheses 2a and 3a**)

	Number of Times Editor i Reverted Others$_{it}$			*Number of Times Editor i Reverted Others Who Reverted*$_{it}$		
Independent Variables	Model 13.	Model 14.	Model 15.	Model 16.	Model 17.	Model 18.
Network Density$_{it-1}$	-.09	-.11	-.05	.15*	.16*	.17*
	(.09)	(.09)	(.09)	(.07)	(.07)	(.07)
Network Size$_{it-1}$	.09***	.02	.09***	.07***	.08***	.08***
	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)
Network Size0$_{it-1}$	.28***	.47***	.34***	-.09	-.19	-.04
	(.09)	(.18)	(.12)	(.20)	(.17)	(.20)
Network Size1$_{it-1}$	.32*	.07	.28*	.27	.26*	.33*
	(.13)	(.13)	(.13)	(.15)	(.12)	(.15)
Percentage of Articles Edited More than Twice$_{it-1}$	-.17	-.08	-.34***	-.11	-.06	-.11
	(.10)	(.11)	(.10)	(.11)	(.11)	(.11)
Cumulative Edits$_{it-1}$	—	.11***	.19***	—	.02	.05
		(.02)	(.02)		(.11)	(.11)
Months since Signup$_{it-1}$	—	-.56***	-.37***	—	.08	-.03
		(.12)	(.11)		(.09)	(.09)
Number of Edits$_{it}$ *x10*	.01***	.01***	-.01***	.47*	.47**	.37
	(.003)	(.003)	(.003)	(.19)	(.16)	(.19)
Time Period Dummies$_t$	No	No	Yes	No	No	Yes
Inverse Mills Ratio$_{it}$	No	No	Yes	No	No	Yes
-Log-Likelihood	25,731	25,401	24,943	20,394	20,073	19,994
Degrees of Freedom	6	8	20	6	8	20
Wald χ^2	12,365	12,466	12,598	12,336	12,367	12,659

Note: Numbers in parentheses are standard errors. Constant was omitted. All χ^2 tests are based on a baseline model with no covariates. Selection equation predicting an undo of a version last saved by editor *i* based on previous experience, tenure and month dummies was omitted. Resulting Inverse Mills ratio was used as control in the outcome equation, as noted. Results of sensitivity tests using generalized linear models with negative binomial link and grouped logits yield equivalent results. Editors i = 30,272; periods t = 12; total number of observations = 212,317 (not all editors started editing in time period 1). As before, results are stable with respect to risk set. *p < .05 **p < .01 ***p < .001 (two-tailed tests).

Table 6. Negative binomial random-effect estimates that editor *i* experienced a revert of an undo at time *t* (Test of **Hypotheses 2b and 3b**)

Independent Variables	*Number of Times Editor i Undone Followed by No Revert$_{it}$*			*Number of Times Editor i Was Undone Followed by Editor i Reverts Undo$_{it}$*			*Number of Times Editor i Was Undone Followed by Another Editor Reverts Undo$_{it}$*		
	Model 19.	Model 20.	Model 21.	Model 22.	Model 23.	Model 24.	Model 25.	Model 26.	Model 27.
Network Density$_{it-1}$	-.21***	.28***	-.17***	-.72***	-.34**	-.93***	.15**	.31**	.19**
	(.04)	(.04)	(.04)	(.14)	(.14)	(.14)	(.06)	(.08)	(.08)
Number of Times Editor i Reverted Others$_{it-1}$	-.10***	-.12***	-.12***	-.25**	-.22**	-.27**	.13**	.18**	.14**
	(.02)	(.02)	(.02)	(.05)	(.05)	(.05)	(.04)	(.04)	(.04)
*Number of Times Editor i Reverted Others$_{it-1}$ * Network Density$_{it-1}$*	-.02**	-.02**	-.02**	-.05**	-.06**	-.05**	.11**	.11**	.12**
	(.005)	(.005)	(.006)	(.01)	(.01)	(.01)	(.03)	(.03)	(.03)
Network Size$_{it-1}$	.61***	.44***	.35***	.45***	.93***	.47***	.73***	.72***	.59***
	(.01)	(.01)	(.01)	(.03)	(.03)	(.03)	(.02)	(.02)	(.02)
Network Size0$_{it-1}$	—	.26***	-.50***	—	-1.49*	-1.80***	—	.55***	.11
	—	(.05)	(.05)	—	(.22)	(.22)	—	(.10)	(.09)
Network Size1$_{it-1}$	—	.46***	-.02	—	-.62	-.95**	—	.75***	.30*
	—	(.07)	(.07)	—	(.32)	(.32)	—	(.15)	(.15)
Number of Articles Edited$_{it-1}$	—	.34***	.25***	—	-.31***	-.44***	—	-.04	-.09***
	—	(.01)	(.01)	—	(.03)	(.03)	—	(.02)	(.02)
Percentage of Articles Edited More than Twice$_{it-1}$	1.08***	1.06***	.99***	.61***	2.22***	1.62***	.98***	.85***	.68***
	(.03)	(.04)	(.04)	(.16)	(.10)	(.10)	(.11)	(.07)	(.07)
Cumulative Edits$_{it-1}$	—	-1.35***	-.46***	—	-.90***	-.21**	—	-.66***	-.80***
	—	(.05)	(.02)	—	(.17)	(.07)	—	(.10)	(.04)
Months since Signup$_{it-1}$	—	1.16***	.28***	—	.77***	.40**	—	.62***	.70***
	—	(.05)	(.02)	—	(.16)	(.06)	—	(.09)	(.04)
Number of Times Editor i Was Undone Followed by No Revert$_{it}$	—	—	—	.03***	.04***	.03***	.04***	.04***	.04***
	—	—	—	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)
Number of Times Editor i Was Undone Followed by Editor i Reverts Undo$_{it}$	.04***	.04***	.03***	—	—	—	.03***	.03***	.00
	(.00)	(.00)	(.00)	—	—	—	(.00)	(.00)	(.00)
Number of Times Editor i Was Undone Followed by Another Editor Reverts Undo$_{it}$	.08***	.07***	.06***	.05***	.11***	.05***	—	—	—
	(.00)	(.00)	(.00)	(.01)	(.01)	(.01)	—	—	—
Number of Times Editor i Undid Others$_{it-1}$	—	—	.31***	—	—	1.16***	—	—	.42***
	—	—	(.01)	—	—	(.02)	—	—	(.01)
Edits$_{it}$x10	.24**	.27*	.15**	.20	.23	.19	.31***	.37	.18
	(.09)	(.10)	(.10)	(.30)	(.33)	(.56)	(.03)	(.32)	(.32)
Time Period Dummies$_t$	No	No	Yes	No	No	Yes	No	No	Yes
Inverse Mills Ratio$_{it}$	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
-Log-Likelihood	110,759	110,547	109,923	20,451	20,211	18,384	36,705	36,409	36,153
Degrees of Freedom	8	14	26	8	14	26	8	14	26
Wald χ^2	30,901	32,173	32,880	6,120	6,454	11,157	9,886	10,332	11,802

Note: Numbers in parentheses are standard errors. Constant was omitted. All χ^2 tests are based on a baseline model with no covariates. Selection equation predicting undos of version saved by that editor was based on previous experience, tenure and month dummies was omitted from table. Resulting Inverse Mills ratio was used as control in the outcome equation, as noted. Results of sensitivity tests using generalized linear models with negative binomial link, and grouped logits yield equivalent results. Editors *i* = 30,272; periods *t* = 12; total number of observations = 212,317. (Not all editors started editing in time period 1.) *p < .05 **p < .01 ***p < .001 (two-tailed tests)

Table 7. Fixed-effects logistic estimates that editor *i* makes at least one edit during time *t* (Test of **Hypotheses 4a-c**)

Independent Variable	Model 28.	Model 29.	Model 30.
Network Density$_{it-1}$	.08*	.08*	.07*
	(.04)	(.04)	(.03)
Number of Times Editor i Was Undone$_{it-1}$	.00		
	(.01)		
Number of Times Editor i Was Undone Followed by No Revert$_{it-1}$		.02	.01
		(.01)	(.01)
Number of Times Editor i Was Undone Followed by Editor i Reverts Undo$_{it-1}$		-.10***	-.07**
		(.02)	(.02)
Number of Times Editor i Was Undone Followed by Another Editor Reverts Undo$_{it-1}$		.08**	.04**
		(.03)	(.01)
Number of Times Editor i Was Undone Followed by Undo Followed by no Revert$_{it-1}$**Density*$_{it-1}$			.08
			(.05)
Number of Times Editor i Was Undone Followed by Editor i Enforces Reverts Undo$_{it-1}$**Density*$_{it-1}$			-.66**
			(.23)
Number of Times Editor i Was Undone Followed by Another Editor Reverts Undo$_{it-1}$ ** Density*$_{it-1}$			.33*
			(.16)
Number of Times Editor i Undid Others$_{it-1}$	Yes	Yes	Yes
Percentage of Articles Editor i Edited More than Twice$_{it-1}$	Yes	Yes	Yes
Network Size$_{it-1}$*, Network Size0*$_{it-1}$ *and Network Size1*$_{it-1}$	Yes	Yes	Yes
Number of Articles Edited$_{it-1}$	Yes	Yes	Yes
Edits$_{it-1}$**10*	Yes	Yes	Yes
Cumulative Edits$_{it-1}$	Yes	Yes	Yes
Months since Signup$_{it-1}$	Yes	Yes	Yes
Periods since Last Edit$_{it-1}$	Yes	Yes	Yes
Time Period Dummies$_t$	Yes	Yes	Yes
-Log-Likelihood	52,421	52,311	51,403
Degrees of Freedom	23	25	28
Wald χ^2	24,188	24,209	24,226

Note: Numbers in parentheses are standard errors. Constant was omitted from table. All χ^2 tests are based on a baseline model with no covariates. Editors *i* = 19,290; number of periods = 12; total number of observations = 153,582. (Not all editors started editing in time period 1.) The number of editors is smaller than in previous tables because fixed-effect estimation removes all editors who always edited or never edited from the risk set. *p < .05 **p < .01 ***p < .001 (two-tailed tests)

References

Ahuja, Gautam. 2000. "Collaboration Networks, Structural Holes and Innovation: A Longitudinal Study." *Administrative Science Quarterly* 45:425-55.

Allison, Paul and Richard Waterman. 2002. "Fixed Effects Negative Binomial Regression Models." *Sociological Methodology* 32:247-65.

Anthony, Denise L, Sean W Smith, and Timothy Williamson. 2009. "Reputation and Reliability in Collective Goods: The Case of the Online Encyclopedia Wikipedia." *Rationality and Society* 21:283-306.

Axelrod, R. 1986. "An Evolutionary Approach to Norms." *American Political Science Review* 80:1095-111.

Barker, James. 1993. "Tightening the Iron Cage: Concertive Control in Self-Managing Team." *Administrative Science Quarterly* 38:408-37.

Bendor, Jonathan and Piotr Swistak. 2001. "The Evolution of Norms." *American Journal of Sociology* 106:1493-545.

Biggart, Nicole. 2001.

Blake, Judith and Kingsley Davis. 1964. "Norms, Values, and Sanctions." in *Handbook of Modern Sociology*, edited by R. Faris. Chicago, IL: Rand McNally.

Buriol, L., C. Castillo, D. Donato, S. Leonardi, and S. Millozzi. 2006. "Temporal Evolution of the Wikigraph." Pp. 45-51 in *Web Intelligence Conference*: IEEE CS Press.

Burt, Ronald S. 1982. *Toward a Structural Theory of Action: Network Models of Social Structure, Perception, and Action*. New York: Academic Press.

—. 2005. *Brokerage and Closure: An Introduction to Social Capital*: Oxford University Press.

Coleman, James. 1989. *American Journal of Sociology*.

Coleman, James Samuel. 1990. *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.

Durkheim, Emile. 1893. *The Division of Labor in Society*.

—. 1951. *Suicide: A Study in Sociology*. Translated by J. A. Spaulding and G. Simpson. New York, NY: The Free Press.

Ellickson, R.C. . 2001. "The Evolution of Social Norms: A Perspective from the Legal Academy." Pp. 35-75 in *Social Norms*, edited by M. Hechter and K.-D. Opp. New York, NY: Russell Sage.

Ellickson, Robert C. 1991. *Order without Law: How Neighbors Settle Disputes*. Cambridge, MA: Harvard University Press.

Elster, Jon. 1989. *The Cement of Society: A Study of Social Order*. New York, NY: Cambridge University Press.

—. 2003. "Coleman on Social Norms." *Revue française de sociologie* 44:297-304.

Fehr, S. and E. Gächter. 2002. "Altruistic Punishment in Humans." *Nature* 415:137-40.

Flache, Andreas and Michael W. Macy. 1996. "The Weakness of Strong Ties: Collective Action Failure in a Highly Cohesive Group." *Journal of Mathematical Sociology* 21:3-28.

Goode, William Josiah. 1978. *The Celebration of Heroes: Prestige as a Social Control System*. Berkeley, CA: University of California Press.

Greif, Avner. 1989. "Reputation and Coalitions in Medieval Trade: Evidence on the Maghribi Traders." *Journal of Economic History* 49:857-82.

Guimaraes, Paulo. 2008. "The Fixed Effects Negative Binomial Model Revisited." *Economics Letters* 99:63-66.

Hausman, J. A., B. H. Hall, and Zvi Griliches. 1984. "Econometric Models for Count Data with an Application to the Patents–R & D Relationship." *Econometrica* 52:909-38.

Hechter, Michael. 1987. *Principles of Group Solidarity*. Berkeley, CA: University of California Press.

Hechter, Michael and Satoshi Kanazawa. 1993. "Group Solidarity and Social Order in Japan." *Journal of Theoretical Politics* 5:455-93.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47:153–61.

Homans, George C. 1950. *The Human Group*. New York, NY: Harcourt, Brace.
Horne, Christine. 2001. "The Enforcement of Norms: Group Cohesion and Meta-Norms." *Social Psychology Quarterly* 64:253-66.
—. 2004. "Collective Benefits, Exchange Interests, and Norm Enforcement." *Social Forces* 82:1037-62.
—. 2007. "Explaining Norm Enforcement." *Rationality and Society* 19:139-70.
Kittur, Aniket, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. "He Says, She Says: Conflict and Coordination in Wikipedia." in *Computer Human Interaction*. San Jose, CA.
Knutson, Brian. 2004. "Sweet Revenge?" *Science* 305:1246-47.
Kossinets, Gueorgi and Duncan J. Watts. 2009. "Origins of Homophily in an Evolving Social Network." *American Journal of Sociology* 115:405-50.
Lin, Nan. 2001. *Social Capital: A Theory of Social Structure and Action* Cambridge University Press.
Menchik, Daniel A. and Xiaoli Tian. 2008. "Putting Social Context into Text: The Semiotics of E-Mail Interaction." *American Journal of Sociology* 114:332-70.
Molm, Linda D. 1997. *Coercive Power in Social Exchange*. Cambridge, UK: University of Cambridge Press.
Morgan, Stephen L. and Aage B. Sørensen. 1999. "Parental Networks, Social Closure, and Mathematics Learning: A Test of Coleman's Social Capital Explanation of School Effects." *American Sociological Review* 64:661-81.
Oliver, Pamela. 1980. "Rewards and Punishments as Selective Incentives for Collective Action: Theoretical Investigations." *American Journal of Sociology* 85:1356-75
Olson, Mancur. 1971. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.
Opp, Karl-Dieter. 1982. "The Evolutionary Emergence of Norms." *British Journal of Social Psychology* 21.
Parsons, Talcott. 1953. "A Revised Analytical Approach to the Theory of Social Stratification." Pp. 99-128 in *Class, Status and Power*, edited by R. Bendix and S. Lipset. New York, NY: Free Press.
Piskorski, Mikolaj Jan. 2010. "Networks as Covers: Evidence from an on-Line Social Network." in *Workin Paper*: Harvard Business School.
Sampson, Robert J., Stephen W. Raudenbush, and Felton Earls. 1997. "Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy." *Science* 277:918-24.
Schotter, Andrew. 1982. *The Economic Theory of Institutions*. Cambridge: Cambridge University Press.
Simmel, Georg. 1902. "The Number of Members as Determining the Sociological Form of the Group. Ii." *American Journal of Sociology* 8:158-96.
Sugden, Robert. 1986. *The Economics of Rights, Co-Operation and Welfare*. Oxford: Basil Blackwell.
Uehara, Edwina. 1990. "Dual Exchange Theory, Social Networks, and Informal Social Support." *American Journal of Sociology* 96:521-57.
Uzzi, Brian. 1999. "Embeddedness in the Making of Financial Capital: How Social Relations and Networks Benefit Firms Seeking Financing." *American Sociological Review* 64:481-505.
Weber, Max. 1976. *The Protestant Ethic and the Spirit of Capitalism*. Translated by T. Parsons and R. H. Tawney. New York, NY: Scribner.
Yamagishi, Toshio and Karen S. Cook. 1993. "Generalized Exchange and Social Dilemmas." *Social Pscyhology Quarterly* 56:235-48.